8-2011

Design and inference in phase II/III clinical trials incorporating monitoring of multiple endpoints.

Herman E. Ray
University of Louisville

DESIGN AND INFERENCE IN PHASE II/III CLINICAL TRIALS INCORPORATING MONITORING OF MULTIPLE ENDPOINTS

By

Herman E. Ray
B.S., Middle Tennessee State University, 2001
M.S., Middle Tennessee State University, 2004

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2011

Design and Inference in Phase II/III Clinical Trials Incorporating Monitoring of Multiple Endpoints

By

Herman E. Ray
B.S., Middle Tennessee State University, 2001
M.S., Middle Tennessee State University, 2004

A Dissertation Approved On

May 20, 2011
Date

By the following Dissertation Committee:

Shesh Rai
Dissertation Director

Somnath Datta

Maiying Kong

Susan Galandiuk

DEDICATION

I dedicate this work to my loving family;

my wife, Jennifer,

my son, Jonathan,

my daughter, Ellie,

and

the "baby-to-be".

This accomplishment is only possible due to the endless love, support, sacrifice, and motivation provided throughout.

ACKNOWLEDGEMENTS

The results contained within these bindings represent hard work and effort, but more importantly relationships, inspiration, motivation, and support. Many people have been involved from the school and my personal life. Without the relationships, I would not have personally grown into the person I have become, nor would this work have been possible to achieve.

The unique opportunities presented by my mentor, Dr. Shesh Rai, created the foundation of the research but also provided situations to develop new skills. Dr. Rai encouraged independent thinking and further exploration of the problems. His remarkable ability to quickly comprehend the completed research and ascertain a deficiency ensured a precise solution. Dr. Rai is a true mentor who has included me in several projects with colleagues from around the campus. He always had faith in my abilities and sincerely trusted my judgments. I started the program hoping to find a mentor, but I leave with a great friend. I will greatly miss the "Thursday" meetings.

Many others contributed in some way to the dissertation. Colleagues and friends from Thomson Reuters, took a risk allowing me to relocate to the Louisville area and work from a home office. Laura Rissover, Brian Hochrien, and Guy Brooks allowed a flexible schedule so that I could still work full-time while also completing this personal goal. This unique opportunity allowed me to study and conduct research while still supporting my family.

The early semesters of the Ph.D. program were extremely challenging, but the guidance and support Dr. Somnath Datta provided is greatly appreciated. His

office was always open. I also learned much about statistics and teaching from his challenging courses.

I am truly grateful for the efforts put forth by the members of the dissertation committee, Dr. Somnath Datta, Dr. Susan Galandiuk, and Dr. Maiying Kong. Reading a large dissertation takes dedication and time.

My parents, Gene and Tammy Ray, always provided endless support, motivation, and love. I appreciate the opportunities and guidance they provided throughout my life. They always supported my dreams and ambitions, regardless of how unachievable they seemed.

Finally, my family sacrificed far more than expected when we made the decision for me to work full-time and go to school. This accomplishment was not possible without their support, motivation, understanding, and unending love. My loving wife, Jennifer, was essentially a single mother of two. My children, Jonathan and Ellie, think the PC is a permanent appendage. Words cannot express my appreciation for all that you endured. This dissertation is complete and I plan to enjoy large quantities of family time.

ABSTRACT

DESIGN AND INFERENCE IN PHASE II/III CLINICAL TRIALS
INCORPORATING MONITORING OF MULTIPLE ENDPOINTS

Herman E. Ray

May 20, 2011

The phase II clinical trial is a critical step in the drug development process. In the oncology setting, phase II studies typically evaluate one primary endpoint, which is efficacy. In practice, a binary measurement representing the response to the new treatment defines the efficacy. The single-arm, multiple-stage designs are popular and the Simon 2-Stage design is preferred.

Although the study designs evaluate the efficacy, the subject's safety is an important concern. Safety is monitored through the number of grade 3 or grade 4 toxic events. The phase II clinical trial design based on the primary endpoint is typically augmented with an ad hoc monitoring rule. The studies are designed in two steps. First, the sample size and critical values are determined based on the primary endpoint. Then an ad hoc toxicity monitoring rule is applied to the study.

Previous authors recommended a method to monitor toxic events after each patient is enrolled which is also known as continuous toxicity monitoring. A trial designed at the JG Brown Cancer Center combined the Simon 2-Stage design with continuous toxicity monitoring. We describe how to integrate the continuous toxicity monitoring methodology with the Simon 2-Stage design for response.

Theoretical justification is given for the nominal size, power, probability of early termination (PET), and average sample size (ASN) of the combined testing procedure. A series of simulations were conducted to investigate the performance of the combined procedure. We discover that the type I error rate, type II error rate, PET, and ASN are subject to the correlation between toxicity and response. In fact, the study may have a smaller type I error rate than expected.

The theoretical expressions derived to describe the operating characteristics of the combined procedure were utilized to create a new flexible, bivariate, multistage clinical trial. The design is considered flexible because it can monitor toxicity on a different schedule than response. An example is considered in which toxicity is measured after four equally spaced intervals and the response is evaluated only at the second and fourth toxicity examinations. This example corresponds to a data monitoring committee's meeting schedule that may happen every 6 months over a two year span. The effect of the correlation on the type I and type II error rates is examined through simulation. The simulations also examine the power over the range of response rates with a fixed toxicity rate in the alternative region and vice-versa.

There are several single-arm, multiple-stage clinical trial designs that consider multiple endpoints at the same time. A subset of the designs includes those that consider both efficacy and toxicity as binary endpoints. A common problem, considered after the conduct of the trial, is appropriate inference given the repeated examinations of the multiple endpoints. We propose a uniformly minimum variance unbiased estimator (UMVUE) for the response in a multistage clinical trial design incorporating toxicity effects. The proposed estimator and the typical maximum likelihood estimator (MLE) are evaluated through simulation. The estimator requires further modification when continuous toxicity monitoring is combined with a multistage design for response. The modified estimator maintains low bias over the range of possible response values.

The larger phase IIb or phase III clinical trial is the logical extension of the

bivariate research based on exact calculations. The phase IIb or III clinical trials typically include an ad hoc toxicity monitoring rule ensuring participant protection. The designs also include provisions to allow early stopping for futility or efficacy utilizing group sequential theory or stochastic curtailment. We also examine a novel large sample clinical trial design that incorporates correlation between the response and toxicity events. The design uses the typical critical values associated with the standard normal distribution. It also searches for critical values specific to the global hypothesis associated with both response and toxicity. The bivariate test is then combined with efficacy and safety monitoring based on a flexible time-varying conditional power methodology. The type I and type II error rates of the bivariate test procedure, along with the bivariate test procedure combined with the conditional power methodology, are investigated through simulation. A modification is developed for the conditional power methodology to preserve the type I and type II error rates.

In the end, the research extends the bivariate clinical trial designs in an attempt to make them more appealing in practice. Although, the research resulted in positive outcomes, additional work is required.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The practice of evaluating treatments in human subjects through a clinical trial is relatively new. Although it can be traced to the 18th century, much of the rigorous statistical work in the area of clinical trials has been conducted within the past 80 years (Friedman et al., 1998). Chow and Liu (2004) observe that many of the statistical advances are directly related to the implementation of regulations designed to protect human subjects that participate in the trials.

The treatments under evaluation progress through phases in the drug development process (Friedman et al., 1998). Traditionally, there are four phases denoted as phase I, II, III, and IV. Each phase has a unique objective that is reflected in the clinical trial design.

A phase I clinical trial is intended to provide the initial evaluation of a new treatment administered to human subjects. The studies attempt to determine metabolic and pharmacological activities of the treatment in humans, side effects of increasing dose, and early evidence of effectiveness (Chow and Liu, 2004). This initial information is required to design the subsequent phase II clinical study. The phase I designs are small, flexible studies that incorporate frequentist or Bayesian principles. In practice, the primary focus of the phase I study is patient safety (Chow et al., 2008) while the phase II trial is concerned with efficacy.

The phase II clinical study is usually designed to test the efficacious attributes of a treatment that passed through a phase I study. The trials are often single-arm studies that test the clinical response rate against some pre-specified value, which represents the maximum response rate that is not clinically interesting

(Stallard et al., 2001). The clinical response rate (referred to as response rate) can consist of complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD) (FDA, 2007). Typically, the studies measure some combination of the different response rates, such as the sum of complete and partial response or objective response (OR). The designs typically incorporate multiple stages, or interim analyses. Green (2006) notes that most phase II trials incorporate two stages and the Simon 2-Stage design (Simon, 1989) is the most popular.

The endpoint utilized in the comparative phase III study may be different than the endpoint used in the phase II study (Friedman et al., 1998). The comparative phase III studies can consider survival endpoints, as well as response rates. The designs often include provisions for early examination of the data using group sequential theory. The early methodologies by Pocock (1977), O'Brien and Fleming (1979), and many others allow early termination of the trial if there is evidence to reject the null hypothesis. Other authors, such as Emerson and Fleming (1989), consider group sequential procedures which allow the trial to terminate early in favor of or to reject the null hypothesis. Jennison and Turnbull (1999) refer to these designs as "inner-wedge" designs since the futility stopping boundaries form a wedge inside of the efficacy boundaries. The designs primarily focus on normally distributed variables, such as clinical response or event free survival.

All clinical trials must also consider the safety of the trial's participants. Chow and Liu (2004) note that the ICH E9 guidelines on statistical considerations in clinical trials stress that safety must be monitored in all clinical trials (ICH, 1999). Therefore, procedures that allow for early termination of the trial for safety reasons should be considered. The patient safety is typically monitored through the number of grade 3 or higher toxicities experienced by the study participants. The toxicity grades are defined by the Common Toxicity Criteria (NCI, 2009). Our research focuses on phase II clinical trials that combine response and safety consideration. A trial designed at the JG Brown Cancer is the motivation of the research.

1.1 Motivating Example

The Simon 2-Stage design (Simon, 1989) was combined with the continuous toxicity monitoring (Ivanova et al., 2005) in an early single-arm phase II trial designed at the JG Brown Cancer Center. Despite the chemotherapy treatment, patients with multiple myeloma tend to relapse due, in part, to drug resistance. Essentially, the damaging effects of the chemotherapy on the myeloma cells can be counteracted by the nurturing bone marrow microenviroment. The ability to inhibit cellular repair would allow apoptosis to continue, thus rendering the treatment more effective. Simvastatin appears to overcome the cell-adhesion mediated drug resistances in ex vivo experiments. It is also known that zoledronic acid increases the effects of simvastatin. The principal investigator hypothesized that simvastatin combined with zoledronic acid will decrease bortezomib and bendamustine drug resistance when treating multiple myeloma patients.

The life expectancy of patients with relapsed or refractory multiple myeloma is not very good. Therefore, the trial must be able to stop early if the combination treatment does not appear to be beneficial so the patients can be moved onto the standard treatment. The Simon 2-Stage design was employed to test the sustained response (SR) rate achieved on the new treatment plan against the corresponding sustained response rate derived from current literature. The SR is defined as either complete response, partial response, or stable disease. The new treatment will be rejected if the SR is 48% (or less) and accepted if it is 68% (or greater). The minimax design was selected with $\alpha_r = 0.05$ and $\beta_r = 0.20$. If there are 11 or more responses in the first 20 patients, then the study will acrue 40 total subjects. If there are 24 or fewer total responses from the 40 subjects, then the treatment will be rejected.

The combination of the two cytotoxic agents could produce a large number of grade 3 or 4 toxic events very quickly. Therefore, we decided to monitor the number of toxic events after each patient is enrolled in the trial. The study could be terminated if there is sufficient evidence that the toxicity rate is greater than or

TABLE 1

Boundaries for Toxicity Monitoring ($P_{T_0} = 0.33$, $\alpha_T = 0.05$, $n = 40$,)

Minimum # Subjects	Maximum # Subjects	# of Subjects with a Toxicity (b_i)
4	4	4
5	6	5
7	7	6
8	9	7
10	11	8
12	14	9
15	16	10
17	18	11
19	19	12
20	22	13
23	23	14
24	27	15
28	28	16
29	30	17
31	32	18
33	34	19
35	37	20
36	40	21

equal to 33%. The probability of stopping early, if the true toxicity rate is less than 33%, was fixed at $\alpha_t = 0.05$. The Pocock boundary (Pocock, 1977) was selected since it affords a greater probability of discontinuing the study early. Table 1 reports the discrete toxicity boundary values. The cumulative number of toxic events after each person is treated will be compared to the boundary values. If the total number of grade 3 or higher toxicities, after person i is treated, is greater than or equal to the associated boundary value, b_i, then the combination treatment is rejected for safety considerations.

The motivating example identified the initial problem to be solved. The Cancer Center required a way to predict the effect of the continuous toxicity monitoring on the Simon 2-Stage design's operating characteristics. The expressions

are crucial to fully understand the effect the correlation has on the combined procedure since there is an underlying assumption of independence between the two endpoints. Chapter 2 contains a review of the current literature associated with the phase II clinical trial designs. Chapter 3 examines the theoretical derivations of the expressions for the operating characteristics of the combined procedure. Chapter 3 also contains a thorough evaluation of the operating characteristics of the combined procedure. Chapter 4 expands the ad hoc design into a formalized phase II clinical trial design that considers response on a different schedule than toxicity. The design includes multiple examinations of response with the ability to continuously monitor toxicity. Inference after the conduct of the bivariate trial is considered in Chapter 5. The concept is expanded into the multiple-arm large sample phase IIb or III setting in Chapter 6. Finally, Chapter 7 contains the concluding remarks and future direction.

CHAPTER 2

REVIEW OF MULTISTAGE CLINICAL TRIALS THAT INCLUDE MULTIPLE ENDPOINTS

The motivating example combined two commonly used methodologies into in one ad hoc procedure. The Simon 2-Stage design (Simon, 1989) was utilized to design the trial based on the response criteria. Group sequential theory, along with exact calculations, were leveraged to determine the continuous toxicity boundary values.

The Simon 2-Stage design is commonly used in the phase II setting. It is based on the binomial distribution and leverages work on phase II clinical trials based on exact calculations by Gehan (1961), Aroian (1968), Schultz, Nichol, Elfring, and Weed (1973), Colton and McPherson (1976), and Fleming (1982). Each author contributed a component to the development of two-stage designs. The result is a popular phase II clinical trial design that controls the type I error rate while allowing an examination of the data before the end of the trial.

The continuous toxicity monitoring methodology relies on group sequential theory, which allows hypothesis testing to be performed after groups of patients are enrolled. The family of group sequential clinical trials are designed in such a way that the type I error rate is preserved while taking advantage of the sequential enrollment of the subjects. The same characteristics are utilized to monitor the cumulative number of toxicities while controlling the type I error, or the probability of declaring a safe treatment unsafe.

The Simon 2-Stage design combined with continuous toxicity monitoring is an ad hoc design. The sample size determination is based only on response, even

though the inclusion of the toxicity monitoring affects the procedure's power to detect meaningful differences. There are phase II clinical trials that examine multiple endpoints simultaneously while allowing multiple examinations of the data.

2.1 Phase II Bivariate Clinical Trial Designs

The phase II clinical trial is often a single-arm multistage clinical trial. The primary end point is typically defined to be a response rate, which is a predefined combination of complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD) (FDA, 2007). The studies are designed to test the response rate of the treatment against some pre-specified value elicited from the principal investigator or literature. The theoretical value represents the response rate under the best treatment currently available.

The safety of the patients is also an important endpoint to consider. The safety is typically monitored through the number of grade 3 or 4 toxic events as defined by the Common Toxicity Criteria (NCI, 2009). There are two general ways to incorporate toxicity into the phase II clinical trial design. One method includes toxicity considerations through formal designs that accommodate multiple endpoints. It is also possible to utilize an ad hoc design in which the toxicity monitoring is developed outside of the design that considers the primary endpoint.

The formal designs include the bivariate two-stage methodologies proposed by Bryant and Day (1995), as well as the multistage bivariate designs proposed by Conaway and Petroni (1995). The two competing designs rely on different theory to develop the procedures and expressions for the operating characteristics. The underlying execution of the designs is the same. First, a pre-defined number of subjects will be enrolled into the trial. If the number of responses is too low, or the number of toxicities is too high, then the trial enrollment is suspended. If not, then another group of subjects is enrolled. In the two-stage setting, if the number of responses is high, and the number of toxicities is low, then the treatment is declared successful. Otherwise, the treatment is declared unsuccessful since the response rate

is too low, or the toxicity rate is too high. Figure 1 displays the hypothesis space. The value, P_{R_0}, specifies the largest response rate that is uninteresting under the null, while P_{R_A} specifies the smallest response rate that is clinically meaningful. The toxicity rate, P_{T_0}, is the smallest toxicity rate that is unacceptable while P_{T_A} is the largest toxicity rate that is acceptable. The stage-wise sample sizes and critical values used to evaluate both response and toxicity are selected to ensure that the specified operating characteristics are met.

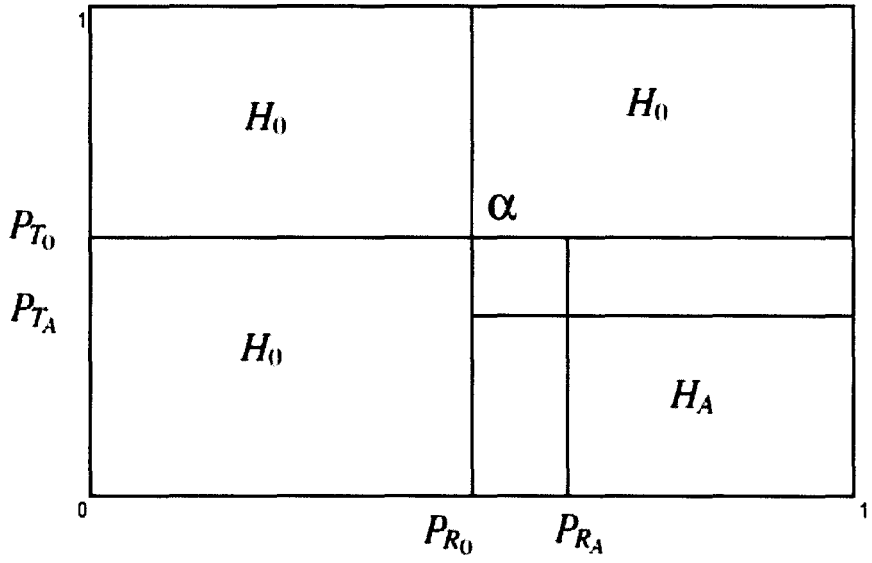

Figure 1. The Hypothesis Space Associated with Conaway and Petroni's Phase II 2-Stage Bivariate Design

The formalized bivariate designs have several issues that limit their utility. Some of the limitations have been addressed while others require additional research.

The methods require specification of the odds-ratio, which relates the response to the toxicity. Tournoux et al. (2007) examine the effect of misspecifying the odds-ratio on the clinical trial. The type I error rate, or the probability of declaring an unsuccessful treatment successful, is affected by incorrect specification

of the odds-ratio. The Bryant and Day design is more robust against the actual odds-ratio being different than specified at the design. It is also possible to adjust the second stage sample size to reflect the odds-ratio based on data accrued in the first stage (Wu and Liu, 2007).

Both designs assume that response and toxicity are equally important to the principal investigator. In reality, the principal investigator may allow for larger toxicity to achieve a larger response rate. Jin (2007) provides a bivariate clinical trial design that incorporates multiple stages, which also allows for tradeoffs between toxicity and response.

The Conaway and Petroni design does not have readily available software to compute the sample sizes and critical values, which limits its utility. The Bryant and Day design does have supporting software, and is a more appealing choice. In practice, it appears that toxicity monitoring rules are developed outside of the clinical design utilized to evaluate the primary endpoint. In other words, the clinical trial is designed in two steps. First, the trial is designed to evaluate the primary endpoint, and then, the toxicity monitoring rule is put into place.

Ivanova et al. (2005) provide an example of a stopping rule for safety that is included in an ad hoc manner. The trial will stop early if 13 or more of the first 20 patients enrolled (n_t) into an arm experienced a toxicity. The toxic event was defined as not being able to tolerate at least 2 courses of treatment. The stopping rule relies on the fact that the

$$P\{(\#\text{ of Toxicities }) \geq 13 \mid P_{T0} = 0.40, n_t = 20\} \leq 0.05.$$

There are instances when the new agent may cause severe toxicities that require continuous monitoring to ensure the participants' safety. Ivanova et al. (2005) provide an expansion of the stopping rule described above that can be applied to situations that require continuous toxicity monitoring. They suggest monitoring the cumulative number of toxic events after each patient is enrolled based on boundary values developed through group sequential theory.

The design is constructed in an ad hoc manner so that the odds-ratio, or the correlation between toxicity and response, is ignored at the design phase of the clinical trial. The total effect on the operating characteristics is also unknown, including the type I and II error rates. Exact formulas for the operating characteristics are vital for the designs to be a possible option.

2.2 Inference After Bivariate Clinical Studies

The principal investigator and others vested in the study require more information than just the determination that the treatment is successful (or not). The ability to create unbiased point estimates and associated confidence intervals after the conduct of a multistage clinical trial that includes multiple endpoints is also important. The estimates are required to evaluate the treatment. Further, these estimates are vital to designing future large phase IIb or III clinical trials.

In the context of continuous variables, Jennison and Turnbull (1999) demonstrate that the maximum likelihood estimate (MLE) is bias due to the multi-modal distribution of the estimate resulting from multistage clinical trials. Jung and Kim (2004) note that the usual MLE utilized after the conduct of the Simon 2-Stage design is also bias due to the optimal sampling effect. The bias is created because we only observe extreme values resulting from crossing either the lower or upper boundary in the early stages. They discover a uniformly minimum variance unbiased estimate (UMVUE) that is applicable to the multistage phase II clinical trials based on exact calculations.

There is only limited research in the area of point estimation following multistage, single-arm phase II clinical trials that incorporate multiple endpoints based on exact calculations. Chang (2009) develops a point estimator that is applicable in the situation but assumes that some of patients are not eligible for all possible responses. In this context, he proposes a MLE, which requires a special numeric algorithm called expectation-maximization (EM) method designed to handle data missing at random.

The point estimator proposed by Chang (2009) is for a specific situation and would likely need modifications to work in the Simon 2-Stage design combined with continuous toxicity monitoring. There is evidence to suggest that the traditional MLE will be biased in this setting as well. There is also some concern that ignoring the toxicity monitoring will cause an issue if the Jung and Kim (2004) point estimator is applied based solely on the response considerations.

2.3 Large Sample Theory - Univariate Designs

The motivating example combined the Simon 2-Stage design with the continuous toxicity monitoring methodology proposed by Ivanova et al. (2005). The continuous toxicity monitoring methodology relies on the group sequential theory that evolved from the truly sequential procedures developed for quality assurance proposes (Jennison and Turnbull, 1999). The group sequential procedures are usually employed in the phase III setting since they require larger sample sizes making normal approximations appropriate.

An important aspect of protecting subjects is the ability to evaluate the data during the conduct of the trial. Patients are typically enrolled into trials sequentially, and the results are available in a similar manner. A naive approach is to repeat the standard significant test through the conduct of the trial. Armitage, McPherson, and Rowe (1969) observe that repeating the usual significant test as subjects are accrued will inflate the type I error rate. The authors also create a numeric integration algorithm for the multivariate normal distribution that is vital to future developments in group sequential theory.

Pocock (1977) proposes a significant improvement over earlier sequential procedures with a closed group sequential procedure. The closed procedure has a maximum sample by which a decision will be made. This is a significant improvement over the open sequential procedures, which would continue to enroll subjects until a decision could be made. Pocock's methodology is based on the normal distribution and the calculations described by Armitage et al. (1969). The

Pocock procedure allocates equal type I error rate to each examination of the data. O'Brien and Fleming (1979) propose a similar design to the Pocock design except that the procedure allocates less type I error early in the conduct of the trial. Lan and DeMets (1983), as well as Slud and Wei (1982), introduce clinical trial designs that improve the methodologies proposed by Pocock, as well as O'Brien and Fleming (1979). A common limitation shared between the Pocock and the O'Brien-Fleming designs is the number and timing of the interim analysis must be determined during the design stage of the trial. Deviation from the initial plan will result in changes in the type II error rate. The Lan and DeMets design "spends" the type I error rate according to an alpha-spending function. The specific timing of the interim analysis does not need to be specified when the trial is designed, nor does the number of interim analysis.

The theory and applications associated with group sequential theory progressed rapidly from the advent of the alpha-spending function. The traditional procedures only stopped early to reject the null hypothesis, but Pampallona and Tsiatis (1994), as well as Emerson and Fleming (1989), formulate "inner-wedge" clinical trial designs. The inner-wedge design allows the trial to stop early to reject or accept the null hypothesis. The designs work with other endpoints such as overall survival leveraging Cox's proportional hazards model (Cox, 1972), or the non-parametric Kaplan-Meier estimator (Kaplan and Meier, 1958). Wei, Su, and Lachin (1990) adapt the group sequential theory so the trials can incorporate repeated measurements on the subjects through generalized estimating equations proposed by Liang and Zeger (1986).

Halperin et al. (1982) propose a method that allows a study to terminate early for futility based on stochastic curtailment, which is a flexible approach to monitoring the emerging results of a clinical trial. Lachin (2005) describes stochastic curtailment as a decision to terminate the trial based on an assessment of the conditional power (CP). CP is the conditional probability that the final result will exceed the critical value given the accrued data along with an assumption about

the data to be observed during the remainder of the study. Ying and Clarke (2010) outline a flexible time-varying conditional power boundary methodology that allocates portions of the type II error over time based on the typical alpha-spending functions. The methodology has similar benefits including the ability to either pre-specify the interim analysis, or use the flexibility associated with the alpha-spending approach to modify the exact timing of the interim analysis. The resulting methodology is an intuitive expansion and application of the conditional power approach to futility monitoring.

2.4 Large Sample Theory - Bivariate Designs

The large phase III clinical trials must also consider the safety of the subjects. The group sequential theory was extended to include multiple endpoints, such as toxicity and response. There are two basic ways multiple endpoints may be included into the design.

The first general method reduces the multiple endpoints into a "global" test statistic such as the Hotelling's t-test (Hotelling, 1931), a χ^2 test, or an F test. Pocock, Geller, and Tsiatis (1987), as well as Tang, Gnecco, and Geller (1989), propose group sequential testing procedures based on O'Brien's generalized least squared statistics (O'Brien, 1984) designed to handle multiple endpoints. Jennison and Turnbull (1991) also developed a group sequential procedure based on exact calculations for the χ^2 and the F statistics.

Although the global test statistic can handle multiple endpoints, in practice it may not be desirable to allow uncontrollable tradeoffs between toxicity and response. It is also possible to develop a test that considers the marginal hypotheses separately but rejection of all individual hypotheses is required to declare the treatment successful. Jennison and Turnbull (1993), as well as Cook and Farewell (1994), propose bivariate test procedures that considers both endpoints separately, but attempts to control the global type I and type II error rates. The group sequential procedures allow one to evaluate both efficacy and futility, as well as

patient safety, through the conduct of the trial. There are also more sophisticated methods proposed to control the type I error rate as well. Chuang-Stein et al. (2007) suggest a method that attempts to control the average type I error rate resulting in slightly different significant levels associated with the marginal hypothesis test.

The procedures proposed by Jennison and Turnbull (1993), Cook and Farewell (1994), Chuang-Stein et al. (2007), and Kordzakhia et al. (2010) create a new multiple comparison issue referred to by Offen et al. (2007) as the reverse multiplicity problem. The typical multiplicity problem arises when a researcher evaluates multiple endpoints, but rejects the null hypothesis if any of the endpoints appear to be significantly different from the control. In this case, the significance levels for testing the individual endpoints are adjusted downward to account for the multiple analyses in order to conserve the overall type I error rate. There are many methods available, such as the Bonferroni correction, the Westfall and Young procedure (Westfall and Young, 1993), or the False Discovery Rate (Benjamini and Hochberg, 1995).

The reverse multiplicity problem appears when we are required to show statistically significant differences on all co-primary endpoints. Depending on the number of co-primary endpoints required, and the correlations between them, the type II error rate of the study could be substantially inflated resulting in much less power than expected (Chuang-Stein et al., 2007). Often, the sample size is increased to control the type II error rate, but Chuang-Stein et al. (2007) demonstrate that the increase in sample size maybe very large. Offen et al. (2007) suggest reducing the multiple endpoints to one (or at least a minimal number).

In Chapter 6, we propose a multiple-stage, group sequential trial that combines the flexible time-varying conditional power methodology proposed by Ying and Clarke (2010) with the customary multiple-endpoints fixed sample size method proposed by Jennison and Turnbull (1993). The result is a flexible multiple-endpoint group sequential procedure that preserves the type I and II error rates.

CHAPTER 3

SIMON'S 2-STAGE DESIGN COMBINED WITH CONTINUOUS TOXICITY MONITORING

The phase II clinical trial plays a critical role in the drug or treatment development process. It is used to ensure the new agent is sufficiently promising to warrant comparison to the current standard treatment in a large phase III study. Typically, the response rate, or the number of positive responses to the treatment, is used to determine if it is sufficiently promising for further investigation. The trials can be designed to accrue data in stages and perform an evaluation of the effectiveness of the treatment after each stage is complete. This design allows the experiment to be stopped early if the therapy does not appear to be beneficial, which prevents the continued administration of an ineffective treatment. Green (2006) observes that most Phase II trials usually employ two-stage designs, and the Simon 2-Stage design (Simon, 1989) is usually preferred.

It is also important, and usually required, to monitor the safety of the patients over the course of the experiment. There are two basic ways to incorporate toxicity considerations into a trial design. The simpler method to implement is an ad hoc rule that stops the trial early if more than a specified number of toxicities occur. The second general methodology relies on theory that uses the bivariate statistic comprised of both the number of responses and the number of toxic events. Conaway and Petroni (1995), as well as Bryant and Day (1995), both create test that reject the agent if the number of responses is too low or the number of toxic events is too high. The Conaway and Petroni design can be extended beyond two stages but still examines the response and toxicity rates at the same time.

Ivanova et al. (2005) expands the ad hoc toxicity monitoring methodology so the toxic events can be monitored after each patient is enrolled, which is also referred to as continuous toxicity monitoring. The continuous toxicity monitoring can be combined with a single or a multiple stage evaluation of response, such as the Simon 2-Stage design. It is uniquely capable of monitoring toxic events when there is a concern the events will be severe while also allowing sufficient time for a measurable response to develop. In this chapter, we derive the operating characteristics of a trial that combines the Simon 2-Stage design with the continuous toxicity monitoring. The characteristics of interest include the probability of early termination, the size of the trial under the null hypothesis, and the average sample size. The motivation of the research is based on the practical application described in Section 1.1.

3.1 Study Design

The Simon 2-Stage design is one of the most popular designs used in phase II clinical trials. There are agents which can produce a large number of toxic events very fast which warrant continuous toxicity monitoring. Ray and Rai (2011a) provide an example of a combination treatment for patients with relapsed, refractory multiple myeloma. The treatment combines simvastatin and zoledronic acid to reduce mediated drug resistances with bortezomib and bendamustine. The combination of the cytotoxic agents may induce a large number of toxic events very quickly. They explore a procedure that combines the Simon 2-Stage design with the continuous toxicity monitoring, which will be referred to as the *combined procedure.* The combined procedure was used in the clinical trial designed to evaluate the new combination treatment.

The combined procedure assumes the i^{th} person can only experience one of the four possible outcomes at one time. The underlying theory assumes that X_{rti}, where r indicates if a response occurred and t indicates if a toxic event occurred, follows a multinomial distribution with with parameters P_{00}, P_{01}, P_{10}, and P_{11}. A

response for the i^{th} person is $X_{Ri} = X_{10i} + X_{11i}$ and a toxic event is $X_{Ti} = X_{01i} + X_{11i}$. The probability of a response is denoted as $P_R = P_{10} + P_{11}$ while the probability of a toxicity is denoted as $P_T = P_{01} + P_{11}$.

The methodology is implemented in two steps. First, the Simon 2-Stage design parameters are required which include specification of the desired type I and II error rates, as well as P_{R_0} and P_{R_A}. The value P_{R_0} is the largest response rate that is not clinically interesting while P_{R_A} is the smallest response rate that is clinically meaningful. Simon's 2-Stage procedure returns the first stage sample size, n_1, the minimum number of responses required to continue onto the second stage, r_1, the total sample size, n, and the total number of responses required to reject the null hypothesis, r.

The next step is to determine the boundaries used to evaluate the cumulative number of toxicities experienced by the patients accrued up to that point. The information required to determine the boundaries include the maximum tolerated toxicity rate, the total sample size, n, and the probability of stopping the trial early if the true toxicity rate is less than or equal to the specified toxicity rate under the null hypothesis. The usual group sequential theory, including the integrals over the multivariate normal distribution, is leveraged to obtain the boundary values, $\{a_1, a_2, \ldots, a_n\}$, associated with the standard normal test statistics. The original algorithms proposed by Armitage et al. (1969), as well as the algorithm proposed by Zhang and Rosenberger (2008), can be utilized to determine the boundary values. The resulting boundary values must be transformed so they can be used against the cumulative number of toxic events experienced after each patient is enrolled. The transformed boundaries will be denoted as $\{b_1, b_2, \ldots, b_n\}$. It should be noted that the actual probability of stopping the trial early under the null hypothesis may be larger than specified after the boundary values are transformed. Jennison and Turnbull (1999) suggest tweaking the resulting boundaries to achieve the desired probability under the null hypothesis.

The trial can be conducted with the response and toxicity boundaries in place. The cumulative number of toxicities experienced after each patient is enrolled, from 1 to $(n_1 - 1)$, is compared to the associated boundary values b_1 through b_{n_1-1}. If the number of toxicities is at or greater than the associated boundary value, then the trial is stopped. Once n_1 patients are enrolled, then the total number of toxicities is evaluated against the boundary b_{n_1} and the total number of responses is compared to the value r_1. The study should be halted if either the number of toxicities is greater than $(b_{n_1} - 1)$ or if the number of responses is less than $(r_1 + 1)$. If the trial continues, then the number of toxic events that occur after patients $(n_1 + 1)$ through $(n - 1)$ are assessed with the associated boundary values b_{n_1+1} through b_{n-1}. If the number of toxic events is equal to or exceeds any of the associated boundary values, then the trial should be terminated. If the trial enrolls all n subjects, then the total number of responses and the total toxicities are evaluated against the response boundary, r, and the toxicity boundary, b_n, respectively. If the number of responses is less than $(r + 1)$ or if the number of toxicities is greater than $(b_n - 1)$, then the null hypothesis is not rejected. The appropriate conclusion is the new agent is not sufficiently promising. If the response is larger than the value r and the number of toxicities is less than the value b_n, then the null hypothesis is rejected. Thus we can conclude that the new agent is sufficiently promising.

3.2 Determine Stopping Boundaries for Continuous Toxicity Monitoring

The Cancer Center required a method to produce the continuous toxicity monitoring methodology. First, we will discuss an algorithm to calculate the toxicity boundary values based on the usual group sequential theory. The general logic can be applied to the exact calculations described by Pocock (1977) or to the alpha-spending approach described by Lan and DeMets (1983). We will discuss the application of the logic to the alpha-spending approach since it allows greater flexibility.

In a manner similar to Jennison and Turnbull (1999), let $\{Z_1, Z_2, \ldots, Z_K\}$ be a sequence of standardized test statistics associated with a group sequential test. Assume the sequence of test statistics have the following three properties

1. $(Z_1, Z_2, \ldots, Z_K)$ is multivariate normal
2. $E(Z_k) = \Theta\sqrt{(I_k)},\ k = 1, 2, \ldots, K$
3. $Cov(Z_i, Z_j) = \sqrt{\frac{I_i}{I_j}},\ 1 \leq i \leq j \leq K$

which is equivalent to saying that the sequence of standardized test statistics have the canonical joint distribution with information levels $\{I_1, I_2, \ldots, I_K\}$ for the parameter Θ. Note that $Cov(Z_i, Z_j) = \sqrt{\frac{n_i}{n_j}},\ 1 \leq i \leq j \leq K$ in the smiple setting. This also implies that the sequence $\{Z_1, Z_2, \ldots, Z_K\}$ is Markov.

The sequence of probabilities

$$P\{Z_1 \geq a_1\} = \alpha(t_1)$$
$$P\{Z_2 \geq a_2, Z_1 < a_1\} = \alpha(t_2) - \alpha(t_1)$$
$$\vdots$$
$$P\{Z_K \geq a_K, Z_1 < a_1, Z_2 < a_2, \ldots, Z_{K-1} < a_{K-1}, \} = \alpha(t_K) - \alpha(t_{K-1}).$$

can be constructed to calculate the boundary values, $\{a_1, \ldots, a_k\}$, where $\alpha(t_i)$ is the alpha-spending function at time t_i. The multi-dimensional integrals are typically evaluated through the recursive formulas using the methods developed by Armitage et al. (1969). The integration method is efficient but also somewhat cumbersome to implement in custom solutions.

Zhang and Rosenberger (2008) present an alternative algorithm that is simpler to implement. Their methodology uses the correlation matrix and the integration methods developed by Genz (1992), which are available in both SAS and R. The correlation matrix is used to define the relationship between successive test statistics. The fact that $R^{-\frac{1}{2}}(Z_1, Z_2)' \sim BVN(0, I)$ is used to establish the required integral where R is the correlation matrix, Z_1 and Z_2 are standardized test statistics, and I is the identity matrix. Then

$$P(Z_1 < a_1, Z_2 \geq a_2) = \frac{1}{2\pi\sqrt{1-\rho}} \int_{a_2}^{\infty} \int_{-\infty}^{a_1} exp\left\{ \frac{-1}{2(1-\rho)} (z_1^2 - 2\rho z_1 z_2 + z_2^2) \right\} \quad (1)$$

,

with $\rho =$ corr(Z_1, Z_2), can be utilized to find the Pocock or the O'Brien-Fleming boundaries. The correlation can be specified to represent equal spacing of the interim analysis which results in boundary values equivalent to those produced by the ld98.exe software (Reboussin et al., 1998).

Now, the total number of toxic events after each patient is enrolled will be compared to a corresponding boundary value. It is important to note that the algorithm produces boundary values, $\{a_1, a_2, \ldots a_N\}$, associated with the normalized test statistics, $(Z_1, Z_2, \ldots, Z_N)$, where N is the total sample size and $n = 1, 2, 3, \ldots, N$ is number of patients accrued. The values $\{a_1, a_2, \ldots, a_N\}$ must be transformed into the integer boundary values, $\{b_1, b_2, \ldots, b_N\}$, for the results to be applied to monitoring of the number observed toxicities. The boundaries, $\{b_1, b_2, \ldots, b_N\}$, are calculated with the following formula

$$b_i = \left\lceil \left\{ a_i \sqrt{\frac{P_{T_0}(1 - P_{T_0})}{i}} + P_{t_0} \right\} i \right\rceil \quad (2)$$

where i represents the i^{th} observation in the sequence $i = 1, 2, \ldots, N$ (Ray and Rai, 2011a). The resulting size of the test using the integer values, $\{b_1, b_2, \ldots, b_N\}$, maybe larger than the desired size of the test. Jennison and Turnbull (1999) suggest modifying the resulting boundary values, $\{b_1, b_2, \ldots, b_N\}$, to achieve the desired characteristics, including the type I error rate. Exact calculations described by Jennison and Turnbull should be used to determine the type I error rate after the transformation. The choice of alpha-spending function will determine the operating characteristics of the trial or control the probability of stopping the trial early (Ivanova et al., 2005).

3.3 Explanation of the Combined Procedure

Suppose that the response and toxicity are both binary outcomes observed for each patient enrolled in the trial. The toxicity is continuously monitored after each patient is enrolled and the response is monitored according to a Simon 2-Stage design. The procedure is designed to stop early if there is significant evidence to accept either null hypothesis H_{10} or H_{20}, which are defined as

$$\begin{aligned} H_{10} : P_R \leq P_{R_0} \quad &\text{versus} \quad H_{1A} : P_R > P_{R_0} \\ &\text{and} \\ H_{20} : P_T \geq P_{T_0} \quad &\text{versus} \quad H_{2A} : P_T < P_{T_0}. \end{aligned} \tag{3}$$

The value P_{R_0} is the largest response rate that is not clinically interesting and P_{T_0} is the maximum tolerated toxicity rate. The Simon 2-Stage parameters will be denoted as r_1, n_1, r, and n.

Let X_{rti} be the observation associated with the i^{th} patient where r indicates if a response occurred and t indicates if a toxic event occurred. Then $X_{rti} = (X_{00i},\ X_{01i},\ X_{10i},\ X_{11i})$ follows a multinomial distribution with parameters $P_{00},\ P_{01},\ P_{10}$, and P_{11}. A response for the i^{th} subject is $X_{Ri} = X_{10i} + X_{11i}$ and a toxic event is $X_{Ti} = X_{01i} + X_{11i}$. The probability of a response is $P_R = P_{10} + P_{11}$ and the probability of a toxicity is $P_T = P_{01} + P_{11}$. The data layout for the i^{th} person is displayed in Table 2. Let $Y_{rtm} = \sum_{i=1}^{m} X_{rti}$ be the accumulated data up to and including the mth observation for $r = 0, 1$ and $t = 0, 1$. The total number of responses will be defined as $Y_{Rm} = \sum_{i=1}^{m} \sum_{t=0}^{1} X_{1ti} = \sum_{i=1}^{m} X_{Ri}$ and the total number of toxicities will be $Y_{Tm} = \sum_{i=1}^{m} \sum_{r=0}^{1} X_{r1i} = \sum_{i=1}^{m} X_{Ti}$. The stages of the trial will be called $m \in \{1, \ldots, n_1, \ldots, n\}$ and include the Simon 2-Stage time periods in the sequence. This serves to simplify the formulas and the understanding. In a manner similar to that described by Jennison and Turnbull (Jennison and Turnbull, 1999), we define the $C_m(Y_{Rm}, Y_{Tm} | P_{00}, P_{01}, P_{10}, P_{11})$ as the probability of reaching stage m with Y_{Rm} responses and Y_{Tm} toxicities given the underlying

TABLE 2

Contingency Table for Response and Toxicity of the *ith* individual

		Toxicity		
		No	Yes	
Response	No	X_{00k}	X_{01k}	
	Yes	X_{10k}	X_{11k}	X_{Rk}
			X_{Tk}	

probabilities P_{00}, P_{01}, P_{10}, P_{11}. We will shorten it to $C_m(Y_R, Y_T|\mathbf{P})$. Then define

$$C_1(Y_R, Y_T|\mathbf{P}) = f(Y_R, Y_T) = P_{00}^{X_{00}} P_{01}^{X_{01}} P_{10}^{X_{10}} P_{11}^{X_{11}}. \tag{4}$$

For $m \leq n_1$, the formula to calculate the probability of stage m is

$$C_m(Y_R, Y_T|\mathbf{P}) = P\{Y_{Rm} = y_{Rm} \cap Y_{Tm} = y_{Tm}|\mathbf{P}\} \tag{5}$$

$$= \sum_{t=max(Y_T-1,0)}^{min(b_{m-1}-1,Y_T)} \sum_{r=max(Y_R-1,0)}^{Y_R} C_{m-1}(r, t|\mathbf{P}) f(Y_R - r, Y_T - t) \tag{6}$$

where b_m is the boundary value associated with the continuous toxicity monitoring defined in Equation (2). The probability of reaching stage $m = (n_1 + 1)$ is

$$C_m(Y_R, Y_T|\mathbf{P}) = \sum_{t=max(Y_T-1,0)}^{min(b_{m-1}-1,Y_T)} \sum_{r=max(Y_R-1,r_1)}^{Y_R} C_{m-1}(r, t|\mathbf{P}) f(Y_R - r, Y_T - t). \tag{7}$$

Then the remaining probabilities $(n_1 + 2)$ through n can be calculated with Equation 5. The probabilities of reaching stage $(n_1 + 1)$ are slightly modified because the response boundaries are also imposed at stage n_1, which reduces the probability of making it to stage $(n_1 + 1)$ under the null hypothesis. In other words, the number of combinations available to progress into the next stage has been reduced by the inclusion of the response and toxicity boundaries.

3.4 Properties of the Combined Procedure

The calculations of the operating characteristics, once the probability of reaching stage m with Y_{Rm} responses and Y_{Tm} toxicities are determined, is similar

to the derivations described by Jennison and Turnbull (1999).

The $C_m(Y_R, Y_T|\mathbf{P})$ can be selected and summed to calculate the size and the probability of early termination (PET) under the null hypothesis. The probabilities of crossing the boundary at any stage $m \neq n_1$ or n is

$$r_m(\mathbf{P}) = \sum_{y=b_m}^{m} \sum_{x=0}^{m} C_m(x, y|\mathbf{P}). \tag{8}$$

At stage n_1

$$r_{n_1}(\mathbf{P}) = \sum_{y=b_{n_1}}^{n_1} \sum_{x=0}^{r_1} C_{n_1}(x, y|\mathbf{P}) \tag{9}$$

and at stage n

$$r_n(\mathbf{P}) = \sum_{y=b_n}^{n} \sum_{x=0}^{r} C_n(x, y|\mathbf{P}). \tag{10}$$

Then the test procedures power function $\pi(\mathbf{P})$ is

$$\pi(\mathbf{P}) = 1 - \sum_{i=1}^{n} r_i(\mathbf{P}). \tag{11}$$

The power function will determine the size of the test under the null hypothesis, and the power to detect differences under the alternative hypothesis. The probability of early termination is

$$PET(\mathbf{P}) = \sum_{i=1}^{n-1} r_i(\mathbf{P}). \tag{12}$$

The expected sample size, or average sample size (ASN), can computed as

$$ASN(\mathbf{P}) = \sum_{i=1}^{n-1} i \cdot r_i(\mathbf{P}) + n(1 - \sum_{i=1}^{n-1} r_i(\mathbf{P})) + 1 \tag{13}$$

The " $+ 1''$ in the formula above is a direct consequence of not being able to stop at the first stage. Thus we are guaranteed to make it through the first stage, or we cannot stop the trial based on the experience of one patient.

3.5 Simulations

A series of simulations was conducted to evaluate the effect of the joint probability on the size and power of the combined procedure. The simulations

included various combinations of P_{R_0}, P_{R_A}, P_{T_0}, α_R, α_T, and β_R, where α_R and β_R are the desired type I and II error rates associated with the Simon 2-Stage design, respectively. The value α_T is the probability of stopping early if the true toxicity rate is less than or equal to P_{T_0}. The specific combinations utilized in the simulations are reported in Table 3. The toxicity boundary values are reported in Table 4. The boundary values reported in Table 4 are after manual tweaking to achieve the desired size. It is also important to note that it is difficult to stop the trial very early due to the large confidence intervals. In many instances, dependent on the selection of the design parameters, it is not possible to stop the trial before 5 patients are accrued.

TABLE 3

Values of the Combined Procedure Parameters Considered for Simulation

Simon 2-Stage Design Parameters								Toxicity Parameters			
P_{R_0}	P_{R_A}	α_R	β_R	r_1	n_1	r	n	P_{T_0}	α_T		
0.30	0.50	0.05	0.10	8	24	24	63	0.33	0.025	0.050	0.100
0.30	0.50	0.10	0.10	7	22	17	46	0.33	0.050	0.100	0.150
0.30	0.50	0.05	0.10	8	24	24	63	0.09	0.025	0.050	0.100
0.30	0.50	0.10	0.10	7	22	17	46	0.09	0.050	0.100	0.150
0.20	0.35	0.05	0.10	8	37	22	83	0.33	0.025	0.050	0.100
0.20	0.35	0.10	0.10	5	27	16	63	0.33	0.050	0.100	0.150
0.20	0.35	0.05	0.10	8	37	22	83	0.09	0.025	0.050	0.100
0.20	0.35	0.10	0.10	5	27	16	63	0.09	0.050	0.100	0.150

TABLE 4

Toxicity Boundary Values

$P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.05$, $\beta_R = 0.10$, $P_{T_0} = 0.33$

α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0.025	# of Toxicities	.	.	.	.	.	.	6	7	8	9	9	10	10	11	11	11	12	12	13	13	14	14	14	15	15	16	16	17	17	18	18
0.050	# of Toxicities	.	.	.	.	.	5	6	7	8	8	8	9	10	10	10	11	11	12	12	12	13	13	14	14	15	15	16	16	17	17	17
0.010	# of Toxicities	.	.	.	.	.	.	.	7	7	8	8	9	9	9	10	11	11	11	11	12	12	13	14	14	14	15	15	15	16	16	17

α_T	# of Observations	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
0.025	# of Toxicities	19	19	19	20	20	21	21	22	22	23	23	23	24	24	25	25	26	26	25	27	28	28	28	28	29	29	30	31	31	31	31	32
0.050	# of Toxicities	18	18	19	19	20	20	20	21	21	22	22	23	23	23	24	24	25	25	25	26	26	27	27	28	28	28	29	29	30	30	31	31
0.100	# of Toxicities	17	18	18	18	19	19	20	20	20	21	21	22	22	23	23	23	24	24	24	25	25	26	26	26	27	27	28	28	28	29	29	30

$P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.10$, $\beta_R = 0.10$, $P_{T_0} = 0.33$

α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0.050	# of Toxicities	.	.	.	.	.	5	6	7	7	8	9	9	9	10	10	11	11	12	12	12	13	14	14
0.100	# of Toxicities	.	.	.	.	.	5	6	7	7	8	8	9	9	9	9	10	11	11	11	12	13	13	13
0.150	# of Toxicities	.	.	.	.	.	5	5	6	6	7	7	8	8	9	9	10	10	10	11	11	12	12	13

α_T	# of Observations	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
0.050	# of Toxicities	14	15	16	16	16	16	17	17	18	18	19	19	19	20	20	21	22	22	22	23	23	24	24
0.100	# of Toxicities	13	14	14	15	15	16	16	16	17	17	18	18	19	19	20	20	20	21	21	21	22	22	23
0.150	# of Toxicities	13	13	14	14	15	15	16	16	16	17	17	18	18	18	19	20	20	20	21	21	21	22	22

$P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.05$, $\beta_R = 0.10$, $P_{T_0} = 0.09$

α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.025	# of Toxicities	.	.	.	.	.	5	5	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	7	8	
0.050	# of Toxicities	.	.	.	.	4	4	4	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7
0.100	# of Toxicities	.	.	.	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	5	6	6	6	6	6	7	7	7

α_T	# of Observations	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
0.025	# of Toxicities	8	8	8	8	8	9	9	9	9	9	10	10	10	10	10	10	11	11	11	11	11	11	11	11	11	12	12	12	12	12	12	12	13
0.050	# of Toxicities	8	8	8	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10	10	10	10	11	11	11	11	11	11	11	12	12	12	12	12
0.100	# of Toxicities	7	7	7	7	8	8	8	8	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10	10	10	10	10	11	11	11	11	11	11

$P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.10$, $\beta_R = 0.10$, $P_{T_0} = 0.09$

α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.050	# of Toxicities	.	.	.	3	3	4	4	4	4	4	5	5	5	5	5	5	5	6	6	6	6	6	6	6	7	7	7	7	7	
0.100	# of Toxicities	.	.	.	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	7	7		
0.150	# of Toxicities	.	.	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	7

α_T	# of Observations	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
0.050	# of Toxicities	7	8	8	8	8	8	8	8	8	9	9	9	9	9	9	10
0.100	# of Toxicities	7	7	7	7	7	8	8	8	8	8	8	8	8	9	9	9
0.150	# of Toxicities	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	8

$P_{R_0} = 0.20$, $P_{R_A} = 0.35$, $\alpha_R = 0.05$, $\beta_R = 0.10$, $P_{T_0} = 0.33$

α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.025	# of Toxicities	.	.	.	.	.	.	.	7	7	8	9	10	10	11	11	11	12	12	13	13	14	14	15	15	15	16	17	17	17	18
0.050	# of Toxicities	.	.	.	.	.	5	6	7	8	8	8	9	9	10	10	11	11	12	12	13	13	14	14	15	15	17	16	16	17	17
0.100	# of Toxicities	.	.	.	.	.	.	6	7	7	8	8	8	9	9	10	11	11	11	12	12	13	13	14	14	15	15	15	15	16	16

α_T	# of Observations	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0.025	# of Toxicities	18	19	19	20	20	20	21	22	22	22	23	23	23	24	24	25	25	26	26	26	27	27	28	28	29	29	29	30	30	31
0.050	# of Toxicities	17	18	18	19	19	20	21	21	21	21	22	22	23	23	23	24	24	25	25	26	26	26	27	27	28	28	29	29	30	30
0.100	# of Toxicities	17	17	18	18	18	19	19	20	20	20	21	21	22	22	23	23	23	24	24	25	25	25	26	26	27	27	27	28	28	29

α_T	# of Observations	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83
0.025	# of Toxicities	31	32	32	32	33	33	34	34	34	35	35	36	36	36	37	37	38	38	39	39	39	40	40
0.050	# of Toxicities	30	31	31	31	32	32	33	33	34	34	34	35	35	35	36	36	37	37	37	38	38	39	39
0.100	# of Toxicities	29	29	30	30	31	31	31	32	32	33	33	33	34	34	35	35	35	36	36	37	37	37	38

$P_{R_0} = 0.20$, $P_{R_A} = 0.35$, $\alpha_R = 0.10$, $\beta_R = 0.10$, $P_{T_0} = 0.33$

α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.050	# of Toxicities	.	.	.	.	.	5	6	7	8	8	8	9	10	10	10	11	11	12	12	12	13	13	14	14	15	15	16	16	17	17
0.100	# of Toxicities	.	.	.	.	.	.	.	7	7	7	8	9	9	9	10	10	11	11	11	12	12	13	14	14	14	15	15	15	16	16
0.150	# of Toxicities	.	.	.	.	4	5	5	6	6	7	7	8	8	9	9	10	10	10	11	11	12	12	13	13	14	14	14	15	15	16

α_T	# of Observations	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
0.050	# of Toxicities	17	18	18	19	19	20	20	20	21	21	22	22	23	23	23	24	24	25	25	25	26	26	27	27	28	28	28	29	29	30	30	31	31
0.100	# of Toxicities	17	17	18	18	18	19	19	20	20	20	21	21	22	22	23	23	23	24	24	24	25	25	26	26	26	27	27	28	28	28	29	29	30
0.150	# of Toxicities	16	16	17	17	18	18	19	19	19	20	20	21	21	21	22	22	23	23	23	24	24	25	25	25	26	26	27	27	27	28	28	28	29

$P_{R_0} = 0.2$, $P_{R_A} = 0.35$, $\alpha_R = 0.05$, $\beta_R = 0.10$, $P_{T_0} = 0.09$

α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.025	# of Toxicities	.	.	.	.	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	8	8	8	8	8
0.050	# of Toxicities	.	.	.	.	4	5	5	5	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7
0.100	# of Toxicities	.	.	2	2	2	3	3	3	3	3	4	4	4	4	5	5	5	5	5	5	5	6	6	6	6	6	6	6	7	7

α_T	# of Observations	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0.025	# of Toxicities	8	8	8	9	9	9	9	9	9	10	10	10	10	10	10	10	11	11	11	11	11	11	12	12	12	12	12	12	12	13
0.050	# of Toxicities	8	8	8	8	8	8	8	8	9	9	9	9	9	10	10	10	10	10	10	10	11	11	11	11	11	11	11	11	12	12
0.100	# of Toxicities	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10	10	10	10	11	11	11	11

α_T	# of Observations	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83
0.025	# of Toxicities	13	13	13	13	13	13	13	13	14	14	14	14	14	14	14	15	15	15	15	15	15	15	15
0.050	# of Toxicities	12	12	12	12	12	13	13	13	13	13	13	13	14	14	14	14	14	14	14	14	15	15	15
0.100	# of Toxicities	11	11	11	11	12	12	12	12	12	12	12	12	13	13	13	13	13	13	13	14	14	14	14

	$P_{R_0} = 0.20$, $P_{R_A} = 0.35$, $\alpha_R = 0.10$, $\beta_R = 0.10$, $P_{T_0} = 0.09$																														
α_T	# of Observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.050	# of Toxicities	.	.	.	.	4	4	4	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7
0.100	# of Toxicities	.	.	.	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	5	6	6	6	6	6	7	7	7
0.150	# of Toxicities	.	.	.	3	3	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6

α_T	# of Observations	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
0.050	# of Toxicities	8	8	8	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10	10	10	10	11	11	11	11	11	11	11	12	12	12	12	12
0.100	# of Toxicities	7	7	7	7	8	8	8	8	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10	10	10	10	10	11	11	11	11	11	11
0.150	# of Toxicities	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9	9	9	9	10	10	10	10	10	10	11	11	11

3.6 Simulation Procedure

The size and power of the various combinations are evaluated in similar manners. The basic concept is described in the following steps which are repeated 100,000 times for each data point displayed in the following graphics.

The simulation is conducted in five basic steps.

- Generate n observations from a correlated bivariate binomial distribution with marginal probabilities $P_R = P(X_{Rk} = 1)$ and $P_T = P(X_{Tk} = 1)$.
- Sequentially monitor the toxicity for subjects 1 to $(n_1 - 1)$. If the cumulative number of toxicities is equal to or exceeds the corresponding boundary, then we stop the trial and accept the null hypothesis.
- If the trial did not terminate, then at observation n_1 sum up the total number of responses. If there are r_1 or fewer responses, then we halt the trial. If there are more than r_1 responses, then compare the number of toxicities through patient n_1 to the associated boundary b_{n_1}. If it is equal to or exceeds the boundary, then discontinue the study.
- If the trial did not stop, then sequentially monitor the toxicity associated with observations $(n_1 + 1)$ to $(n - 1)$. If the number of toxicities is equal to or exceeds any of the associated boundaries for the corresponding number of subjects, then stop the trial.
- Finally, at observation n, we do not reject the null hypothesis if there are r or fewer responses or if the number of toxicities exceeds b_n.

The simulated type I error rate is calculated for each combination of P_{R_0}, P_{R_A}, P_{T_0}, α_R, α_T, β_R reported in Table 5 and $m \cdot P(X_{Rk} = 1, X_{Tk} = 1) = m \cdot P_{11}$ where $m \in \{0, 0.01, 0.02, \ldots, 0.98, 0.99, 1.00\}$. The sequence resulting from the product of m and P_{11} produces the smooth figures discussed in future paragraphs. The power analysis is performed in a two step process. The first step is to

determine all combinations of probabilities associated with response and toxicity. This is completed by creating two sequences from 0 to 1 by 0.01 increments. Then all possible combinations of the two sequences are determined, which allows the complete power surface to be specified. The power surface is determined under the assumption that toxicity and response are independent. Finally, the effect of the joint probability on the power is analyzed. In order to limit the number of possibilities, the response rate is fixed at the alternative value, P_{R_A}, specified in the Simon 2-Stage design. Then the largest toxicity rate which maintains approximately 90% power in the combined procedure is selected. If it is not possible to obtain 90% power, then the largest toxicity rate that achieves the largest power is selected.

3.7 Simulation Results - Nominal Size

Figure 2 graphically displays the alpha-level of the combined procedure when $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.05$, $\beta_R = 0.1$, $P_{T_0} = 0.33$ and $\alpha_T \in \{0.025, 0.05, 0.10\}$. We can see that in each case the size of the resulting test does not exceed the alpha-level specified for efficacy in the Simon 2-Stage design. The nominal size of the bivariate design decreases as the probability of experiencing both events at the same time increases. Larger values of α_T result in a design that becomes more conservative as the probability of a person experiencing both a response and toxicity increases.

Another important observation is the effect of sample size on the nominal size of the combined procedure. Figure 3 displays the type I error rate when $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.10$, $\beta_R = 0.10$, $P_{T_0} = 0.33$ and $\alpha_T \in \{0.05, 0.10, 0.15\}$. The total sample size is 46, which is less than 63 in the previous example. We can see that the nominal size displays the same characteristics except that it does not obtain the desired value when the probability of experiencing both events is 0. The simulation was designed to include Simon 2-Stage design parameters that would result in various sample sizes. A larger total sample size is required when $P_{R_0} = 0.2$, $P_{R_A} = 0.35$, $\alpha_R = 0.10$, and $\beta_R = 0.10$. The required sample size is 63 units and the

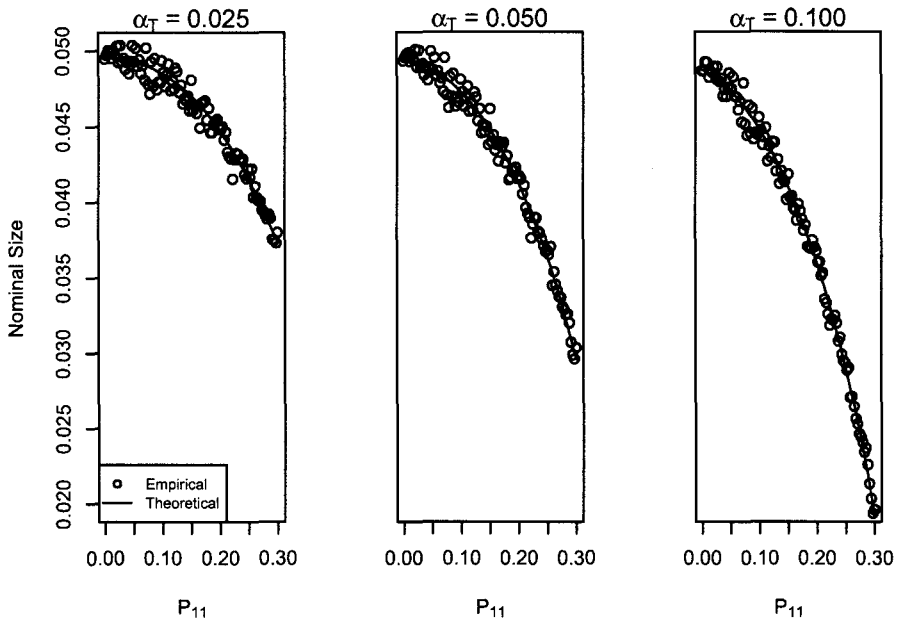

Figure 2. Type I Error of Combined Procedure with $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.05$, and $P_{T_0} = 0.33$

desired alpha-level is almost achieved when the joint probability is 0. Again, larger values of α_T result in a bivariate design that becomes more conservative as the joint probability increases.

3.8 Simulation Results - Power

The surface in Figure 4 is the power surface for all possible combinations of P_{R_A} and P_{T_A} when the design parameters are $P_{R_0} = 0.3$, $P_{T_0} = 0.33$, $\alpha_T = 0.05$, $\alpha_R = 0.05$, and $\beta_R = 0.1$. The surface implies that the procedure successfully identifies agents with a high response rate and low toxicity rate. Figure 5 displays a slice of the power curve over the various toxicity rates when the response rate is fixed at $P_{R_A} = 0.5$. The figure displays the three different curves associated with the different choices of α_T evaluated in the simulation. We can see that smaller values of α_T result in a slightly more powerful combined procedure. The same is true when the alpha-level associated with the Simon 2-Stage design is increased to 0.10 which

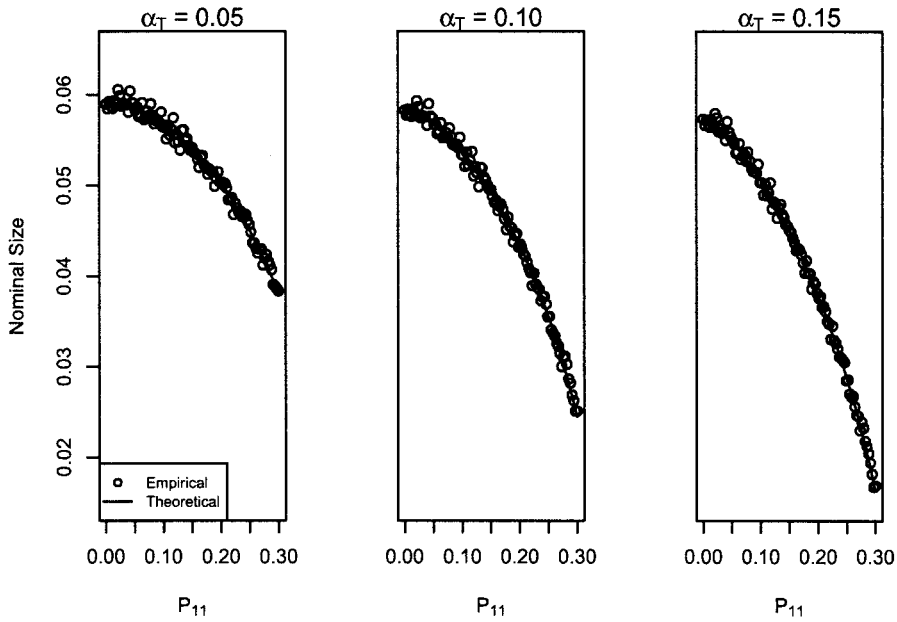

Figure 3. Type I Error of Combined Procedure with $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.10$, and $P_{T_0} = 0.33$

is depicted in Figure 6.

Finally, we examine the effect of the joint probability of a response and toxicity occurring at the same time on the power. In order to examine the effect, we fix the response rate at the alternative value specified in the Simon 2-Stage design. The toxicity rate was selected from the simulated results so the combined procedure's power is close to 0.90. The exact value of 0.90 is not possible to find since we are using simulated results and the underlying distribution is discrete. It is also important to note that the combined procedure may not achieve the power specified in the Simon 2-Stage design. In Figure 6, we see that the maximum power achieved is only 88%. In Figures 7 and 8, we also see that the power declines slightly as the joint probability of experiencing both a response and toxic event increases.

The remaining simulations produce very similar results so they are not included in the discussion. Table 5 contains a summary of the simulation results associated with the nominal size. It is important to note that the combined procedure never achieves the specified alpha-level due to the discrete nature of the

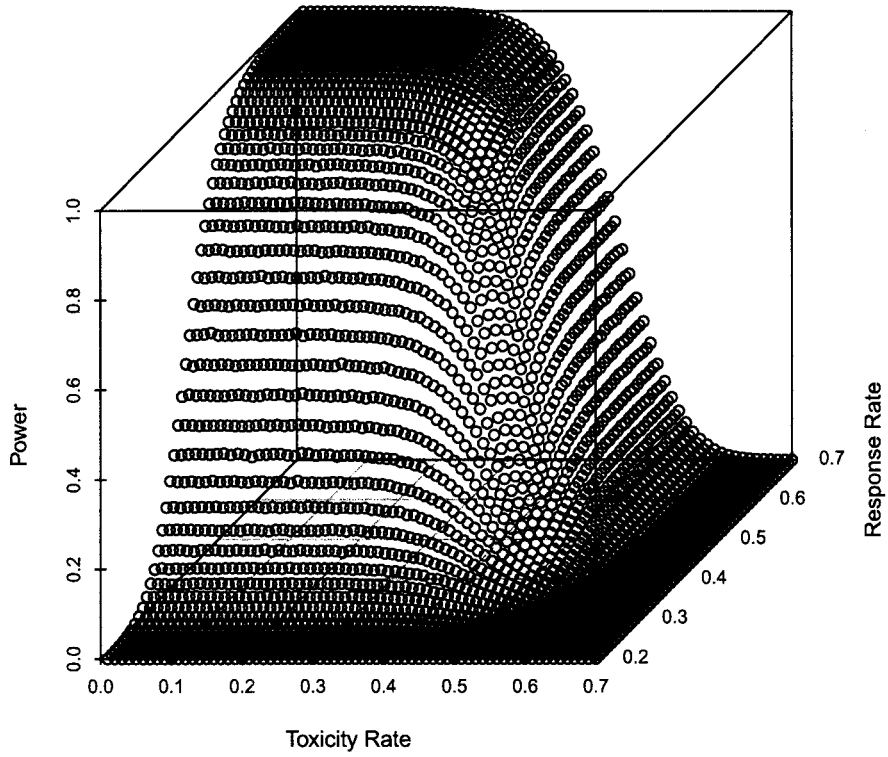

Figure 4. Power Surface of Combined Procedure Over Toxicity Rate with $P_{R_0} = 0.3$, $\alpha_R = 0.05$ and $\alpha_T = 0.05$

distribution. Also, the solid lines in Figures 2, 3, 7, and 8 are graphic representation of the theoretical formulas describing the characteristics of the combined procedure.

3.9 Discussion

The combined procedure is a novel way to include toxicity considerations into a phase II trial designed with the Simon 2-Stage methodology. The result is a trial that incorporates two endpoints simultaneously that does not inflate the type I error rate specified for the response. The boundaries associated with the toxic events can be found using the well established group sequential theory. The

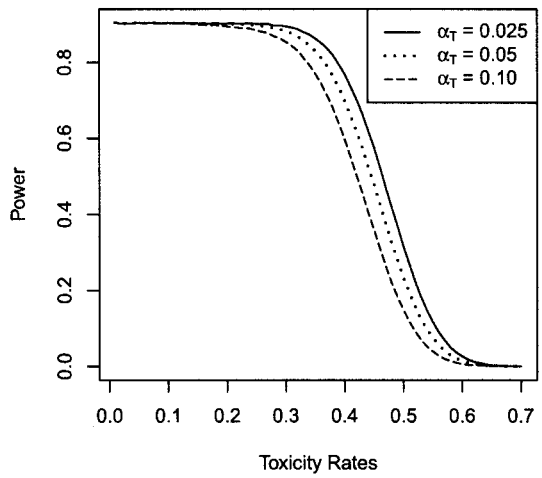

Figure 5. Power Curve of Combined Procedure Over Toxicity Rate with $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, and $\alpha_R = 0.05$

continuous toxicity monitoring methodology requires specification of the maximum tolerated toxicity rate, the probability of stopping the trial early if the toxicity rate is in the null, and the total sample size returned by Simon's procedure.

The simulations indicate that the combination design maintains many of the properties of the Simon 2-Stage design. The overall type I error rate does not exceed the value associated with the response. In fact, the procedure becomes more conservative as the probability of experiencing both a toxicity and a response increases. The sample size may adversely affect both the size and the power of the combined procedure. The desired alpha-level may not be maintained when the joint probability is 0 for small sample sizes. A sample size that is too small may also result in a less powerful combined procedure than desired. Interestingly, the joint probability of experiencing both events at the same time has very little impact on the power, as displayed in figured 7 and 8.

Although the combined procedure has some nice properties, there are some limitations that may reduce its utility. The first limitation is associate patient enrollment. The trial must stop accruing patients while the i^{th} subject is observed.

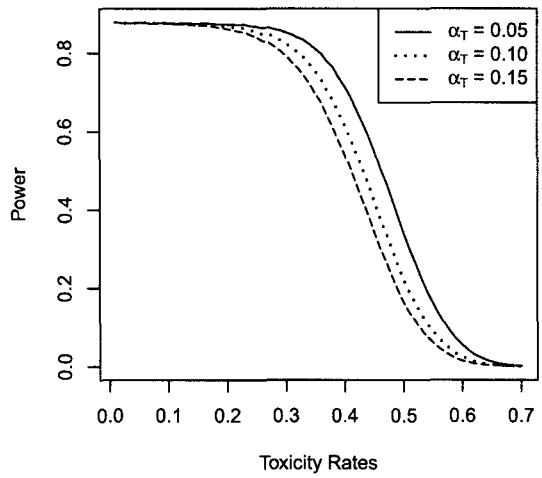

Figure 6. Power Curve of Combined Procedure Over Toxicity Rate with $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, and $\alpha_R = 0.10$

This implies that the trial is beneficial for agents which may cause severe toxicities very quickly. The second limitation is associated with the way the trial is designed. The alternative value is specified for the response but not the toxicity rate. The theoretical expressions, or the simulations, can be utilized to determine the value of the toxicity rate the combined procedure has power to detect, given the alternative value associated with the response. In conclusion, the combined procedure is a viable choice in specific situations since it maintains some of the properties associated with the Simon 2-Stage design. The combined procedure is a feasible way to incorporate toxicity considerations into a single-arm phase II clinical trial.

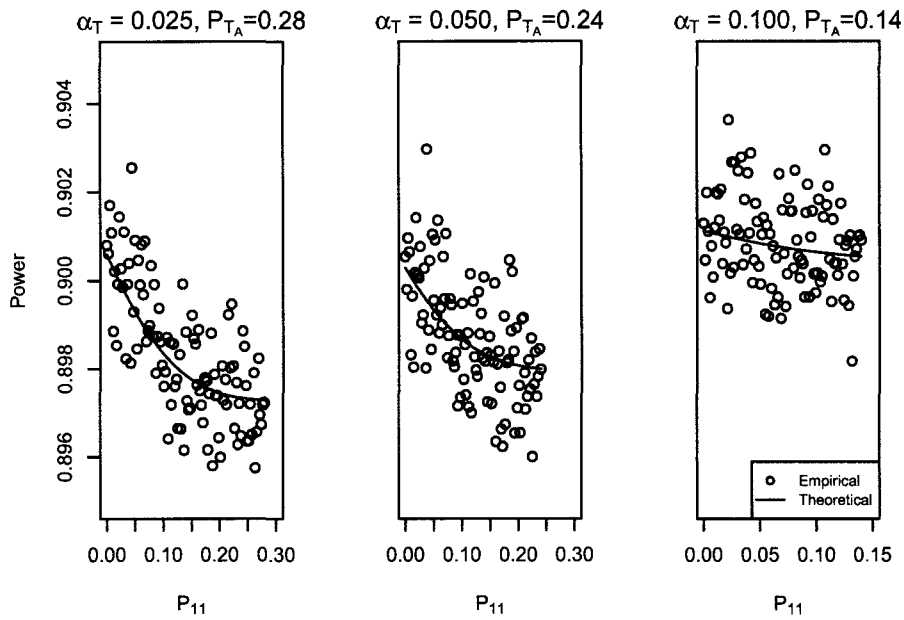

Figure 7. Effect of Joint Probability on Power Curve $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.05$, $\beta_R = .9$, $P_{T_0} = 0.33$. P_{T_A} is the largest alternative toxicity rate which achieves the largest power.

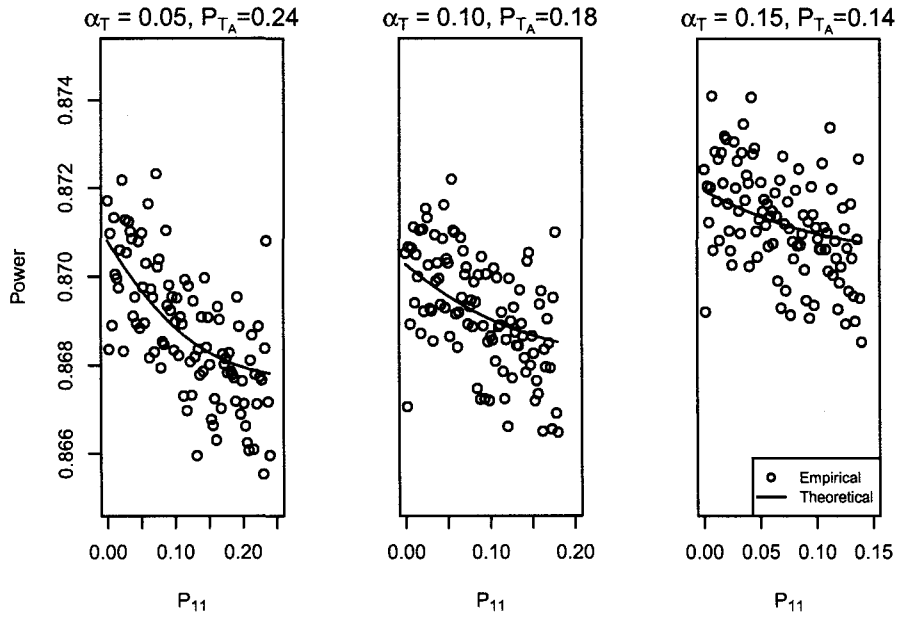

Figure 8. Effect of Joint Probability on Power Curve $P_{R_0} = 0.3$, $P_{R_A} = 0.5$, $\alpha_R = 0.10$, $\beta_R = .9$, $P_{T_0} = 0.33$. P_{T_A} is the largest alternative toxicity rate which achieves the largest power.

TABLE 5

Empirical Size of the Combined Procedure Based on 100,000 Simulations

						$P(X_{Rk}=1$ and $X_{Tk}=1)$ =			
P_{R_0}	P_{R_A}	α_R	β_R	P_{T_0}	α_T	0	$P_R \cdot P_T$	$\min(P_R,\ P_T)$	Actual α_R
0.30	0.50	0.05	0.10	0.33	0.025	0.0495	0.0493	0.0380	0.0497
0.30	0.50	0.05	0.10	0.33	0.050	0.0493	0.0481	0.0304	0.0497
0.30	0.50	0.05	0.10	0.33	0.100	0.0486	0.0456	0.0196	0.0497
0.30	0.50	0.10	0.10	0.33	0.050	0.0590	0.0565	0.0384	0.0974
0.30	0.50	0.10	0.10	0.33	0.100	0.0583	0.0534	0.0251	0.0974
0.30	0.50	0.10	0.10	0.33	0.150	0.0573	0.0504	0.0169	0.0974
0.30	0.50	0.05	0.10	0.09	0.025	0.0491	0.0473	0.0440	0.0497
0.30	0.50	0.05	0.10	0.09	0.050	0.0486	0.0461	0.0408	0.0497
0.30	0.50	0.05	0.10	0.09	0.100	0.0473	0.0437	0.0343	0.0497
0.30	0.50	0.10	0.10	0.09	0.050	0.0588	0.0567	0.0480	0.0974
0.30	0.50	0.10	0.10	0.09	0.100	0.0573	0.0537	0.0413	0.0974
0.30	0.50	0.10	0.10	0.09	0.150	0.0552	0.0505	0.0357	0.0974
0.20	0.35	0.05	0.10	0.33	0.025	0.0474	0.0476	0.0404	0.0487
0.20	0.35	0.05	0.10	0.33	0.050	0.0468	0.0464	0.0356	0.0487
0.20	0.35	0.05	0.10	0.33	0.100	0.0460	0.0440	0.0284	0.0487
0.20	0.35	0.10	0.10	0.33	0.050	0.0986	0.0963	0.0801	0.0999
0.20	0.35	0.10	0.10	0.33	0.100	0.0962	0.0913	0.0667	0.0999
0.20	0.35	0.10	0.10	0.33	0.150	0.0941	0.0862	0.0555	0.0999
0.20	0.35	0.05	0.10	0.09	0.025	0.0485	0.0483	0.0400	0.0487
0.20	0.35	0.05	0.10	0.09	0.050	0.0479	0.0471	0.0352	0.0487
0.20	0.35	0.05	0.10	0.09	0.100	0.0463	0.0441	0.0285	0.0487
0.20	0.35	0.10	0.10	0.09	0.050	0.0981	0.0964	0.0789	0.0999
0.20	0.35	0.10	0.10	0.09	0.100	0.0947	0.0915	0.0656	0.0999
0.20	0.35	0.10	0.10	0.09	0.150	0.0911	0.0865	0.0560	0.0999

CHAPTER 4

FORMALZIED PHASE II BIVARIATE MULTISTAGE DESIGN

The phase II clinical trial is usually designed to test the efficacious attributes of an agent that passed through a phase I study. The trials are often single-arm studies designed to test the clinical response rate against some pre-specified value, which represents the maximum response rate that is not clinically interesting (Stallard et al., 2001). The clinical response rate (referred to as response rate) can consist of complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD). Typically, the studies are designed to measure some combination of the different response rates, such as objective response (OR = CR + PR) or sustained response (SR = CR + PR + SD) (FDA, 2007). The designs can incorporate multiple stages or interim analysis. Green (2006) notes that most phase II trials incorporate two stages with the Simon 2-Stage design (Simon, 1989) being the most popular.

The trials must also incorporate patient safety. In fact, the data monitoring committee is responsible for the protection of the trial participants from harm due to the treatment under study (Demets et al., 2005). The maximum tolerated dose may not be clearly defined by the small phase I study, and thus might lead to more toxic events than desired (Tournoux et al., 2007). There are formal two-stage designs that incorporate both response and toxicity. Typically, grade 3 and above toxic events are monitored where the Common Toxicity Criteria (NCI, 2009) is used to assess the level of toxicity. The earliest such designs include both frequentist and Bayesian philosophies. The first frequentist based designs, which include both

response and toxicity, were proposed by Conaway and Petroni (1995), as well as Bryant and Day (1995). Both designs rely on exact calculations and maintain the small sample sizes expected in phase II trials. The concept was recently expanded by Jin (Jin, 2007) to allow for tradeoffs between toxicity and response to be included in the trial design. Thall, Simon, and Estey (Thall et al., 1996) provide a Bayesian design that sequentially monitors safety and efficacy at the same time.

Often, the univariate response statistic is augmented with an ad hoc toxicity stopping rule. The toxicity monitoring assists the data monitoring committee with the decision to continue the study while protecting the trial participants. The ad hoc stopping rules typically examine the data once during the trial. Ivanova et al. (2005) provide an example of a stopping rule for safety that is included in an ad hoc manner. The trial will stop early if 13 or more of the first 20 patients enrolled into an arm experienced toxicity. In some instances, this may not provide enough guidance. So, they propose a toxicity monitoring schedule that examines the number of toxicities after each subject is enrolled in the trial. Ray and Rai (2011a) combine the continuous toxicity monitoring with the Simon 2-Stage design (Simon, 1989). They discover theoretical expressions that accurately reflect the operating characteristics of the combined procedure. They also note that the continuous toxicity monitoring may be too aggressive in some instances since the trial will need to stop after each patient is treated.

In this chapter, we propose a toxicity monitoring schedule that is combined with a multistage design for response. The toxicity monitoring can be performed on groups of patients. This procedure can also be applied more frequently than the evaluation for response. For instance, the toxicity monitoring can be performed four times, while the response evaluation is performed only twice. The safety examination can correspond with the number and time of the data monitoring committee's meetings, so they have statistical guidance for the decisions required of them. The design includes both endpoints of interest, leverages exact calculations, and maintains the expected sample sizes in phase II trials.

4.1 Historical Designs

The clinical trial design evaluates both response and toxicity simultaneously. The specific hypotheses the design evaluates are

$$\begin{aligned} H_0 &: P_R \leq P_{R_0} \quad \text{or } P_T \geq P_{T_0} \\ &\text{versus} \\ H_1 &: P_R > P_{R_A} \quad \text{and } P_T < P_{T_A}. \end{aligned} \tag{14}$$

where P_{R_0} is maximum response that is not clinically interesting, P_{R_A} is the minimum response rate that is clinically interesting, P_{T_0} is the maximum tolerated toxicity rate, and P_{T_A} is the minimum toxicity that is unacceptable.

In order to find the sample size we also need to specify the global type I error rate, α, as well as the type II error rate, β. We also need α_T which is the probability of declaring a non-toxic treatment toxic based solely on the marginal count of the toxic events. In terms of our hypothesis, the type I error is declaring a new agent sufficiently promising when the response rate is too low, the toxicity rate is too high, or both. The type II error is the probability of rejecting a sufficiently promising agent.

Suppose that both response and toxic events are binary outcomes of a single-arm phase II trial. Let X_{rti} be the observation of the i^{th} person with response r and toxicity t $(r, t = 0, 1)$. Then $X_{rti} = (X_{00i}, X_{01i}, X_{10i}, X_{11i})$ follows a multinomial distribution with parameters P_{00}, P_{01}, P_{10}, and P_{11}. A response for the i^{th} person is $X_{Ri} = X_{10i} + X_{11i}$ and a toxic event is $X_{Ti} = X_{01i} + X_{11i}$. The probability of a response is $P_R = P_{10} + P_{11}$ and the probability of a toxicity is $P_T = P_{01} + P_{11}$. The data layout for the i^{th} person is displayed in Table 6. Let $Y_{rtm} = \sum_{i=1}^{m} X_{rti}$ be the accumulated data up to and including the m^{th} observation for $r = 0, 1$ and $t = 0, 1$. The total number of responses will be defined as $Y_{Rm} = \sum_{i=1}^{m} \sum_{t=0}^{1} X_{1ti} = \sum_{i=1}^{m} X_{Ri}$ and the total number of toxicities will be $Y_{Tm} = \sum_{i=1}^{m} \sum_{r=0}^{1} X_{r1i} = \sum_{i=1}^{m} X_{Ti}$.

TABLE 6

Contingency Table for Response and Toxicity of the i^{th} individual

		Toxicity		
		No	Yes	
Response	No	X_{00i}	X_{01i}	
	Yes	X_{10i}	X_{11i}	X_{Ri}
			X_{Ti}	

The underlying theory assumes that the four possible responses follow a multinomial distribution. Aitken and Gonin (1935) provide the details that relate P_T and P_R through the correlation

$$\rho = \frac{P_{11} - P_R P_T}{\sqrt{P_R(1-P_R)P_T(1-P_T)}}$$

where P_R and P_T could fall in either the null or alternative regions. We can see that ρ is an increasing function of the joint probability P_{11}, and that it will not take on all possible values in [-1,1]. The effect of the choice of ρ will be examined in the simulations.

The Conaway and Petroni bivariate test procedure is designed to evaluate both response and toxicity at the same times. Let Y_{Tn_k} be the total number of toxicities experienced in the first n_k subjects and Y_{Rn_k} be the corresponding number of responses in the first n_k patients, for k in $\{1, \ldots, K-1\}$. The total number of responses and toxicities through all n subjects are denoted by Y_{Rn} and Y_{Tn}, respectively. Let the critical region be denoted as

$$\begin{aligned} CR \;=\; & \{(Y_{Tn_1}, Y_{Tn_2}, \ldots, Y_{Tn_{K-1}}, Y_{Tn}; Y_{Rn_1}, Y_{Rn_2}, \ldots, Y_{Rn_{K-1}}, Y_{Rn}) : \\ & Y_{Tn_1} < T_1, Y_{Tn_2} < T_2, \ldots, Y_{Tn_{K-1}} < T_{K-1}, Y_{Tn} < T; \\ & Y_{Rn_1} > R_1, Y_{Rn_2} > R_2, \ldots, Y_{Rn_{K-1}} > R_{K-1}, Y_{Rn} > R\} \end{aligned}$$

where T_1, T_2, ..., T_{K-1}, T is the toxicity boundaries used at each evaluation and R_1, R_2, ..., R_{K-1}, R is the set of response boundaries assuming K stages. Then they

design the trial while controlling the following error rates

$$\begin{aligned} P\{(Y_{Tn_1}, Y_{Tn_2}, \ldots, Y_{Tn_{K-1}}, Y_{Tn}; Y_{Rn_1}, Y_{Rn_2}, \ldots, Y_{Rn_{K-1}}, Y_{Rn}) \in CR| \\ (P_{R_0}, P_{T_0}), \rho\} \leq \alpha \\ \sup_{H_0} P\{(Y_{Tn_1}, Y_{Tn_2}, \ldots, Y_{Tn_{K-1}}, Y_{Tn}; Y_{Rn_1}, Y_{Rn_2}, \ldots, Y_{Rn_{K-1}}, Y_{Rn}) \in CR| \\ (P_R, P_T), \rho\} \leq \gamma \\ P\{(Y_{Tn_1}, Y_{Tn_2}, \ldots, Y_{Tn_{K-1}}, Y_{Tn}; Y_{Rn_1}, Y_{Rn_2} \ldots, Y_{Rn_{K-1}}, Y_{Rn}) \in CR| \\ (P_{R_A}, P_{T_A}), \rho\} \geq 1 - \beta \end{aligned}$$

Ray and Rai (2011a) considered the situation in which the toxicity is measured after each individual but the response is monitored based on a Simon 2-Stage schedule. First, design the trial for the efficacious endpoint based on the Simon 2-Stage design. Then the toxicity boundaries are constructed using group sequential theory which requires the minimum toxicity rate considered unacceptable, the total sample size, and probability of stopping early given a safe treatment, α_T. The corresponding continuation region, CR, is different and defined to be

$$\begin{aligned} CR_{ctm} = \{(Y_{T1},\ Y_{T2},\ \ldots,\ Y_{Tn};\ Y_{Rn_1},\ Y_{Rn}) : \\ Y_{T1} < T_1,\ Y_{T2} < T_2,\ \ldots, Y_{Tn} < T; Y_{Rn_1} > R_1,\ Y_{Rn} > R\} \end{aligned}$$

Then the trial is designed with the following constraints

$$\begin{aligned} P\{Y_{Rn_1} > R_1, Y_{Rn} > R|P_{R_0}\} &\leq \alpha_R \\ P\{Y_{Rn_1} > R_1, Y_{Rn} > R|P_{R_A}\} &\geq 1 - \beta_R \\ \sum_{i=1}^{n} \alpha_{T_i} &\leq \alpha_T \end{aligned}$$

where the toxicity boundary values are determined as

$$\begin{aligned} P\{Y_{T1} \geq T_1|P_{T0}\} &= \alpha_{T_1} \\ P\{Y_{T2} \geq T_2, Y_{T1} < T_1|P_{T0}\} &= \alpha_{T_2} - \alpha_{T_1} \\ &\vdots \\ P\{Y_{Tn} \geq T, Y_{T1} < T_1, Y_{T2} < T_2, \ldots, Y_{T(n-1)} < T_{n-1}, |P_{T0}\} &= \alpha_{T_n} - \alpha_{T_{n-1}} \end{aligned}$$

The values α_R and β_R are required to construct to the Simon 2-Stage design. The value α_T is the type I error rate associated with the toxicity monitoring. The investigation of the design led to theoretical expressions that can accommodate the evaluation of two different endpoints on two different schedules. Now, utilizing the theoretical expressions we construct a bivariate test which also incorporates the joint probability of both events in order to control the type I and type II error rates associated with the bivariate hypothesis.

4.2 New Design

Let K be the planned number of toxicity evaluations and L be the planned number of response evaluations with $K \neq L$. The continuation region is defined to be

$$\begin{aligned} CR_{RR} = \{&(Y_{Tn_1}, Y_{Tn_2}, \ldots, Y_{Tn_k}, \ldots, Y_{Tn_{K-1}}, Y_{Tn}; \\ &Y_{Rn_1}, Y_{Rn_2}, \ldots, Y_{Rn_l}, \ldots, Y_{Rn_{L-1}}, Y_{Rn}) : \\ &Y_{Tn_k} < T_k, k \in \{1, 2, \ldots (K-1)\}, \\ &T_{Rn_l} > R_l, l \in \{1, 2, \ldots, (L-1)\},\ k = l \nRightarrow n_k = n_l, \\ &Y_{Tn} < T, Y_{Rn} > R\}. \end{aligned}$$

The trial is constructed with the following constraints

$$\begin{aligned} &P\{(Y_{Tn_1},\ Y_{Tn_2} \ldots, Y_{Tn_{K-1}},\ Y_{Tn}; \\ &\qquad Y_{Rn_1},\ Y_{Rn_2}, \ldots, Y_{Rn_{L-1}},\ Y_{Rn}) \in CR_{RR} | (P_{R_0}, P_{T_0}), \rho\} \leq \alpha \\ &P\{(Y_{Tn_1},\ Y_{Tn_2} \ldots, Y_{Tn_{K-1}},\ Y_{Tn}; \\ &\qquad Y_{Rn_1},\ Y_{Rn_2}, \ldots, Y_{Rn_{L-1}},\ Y_{Rn}) \in CR_{RR} | (P_{R_A}, P_{T_A}), \rho\} \geq 1 - \beta \\ &\qquad \sum_{k=1}^{K} \alpha_{T_k} \leq \alpha_T \end{aligned}$$

where the toxicity boundary values are determined as

$$\begin{aligned} P\{Y_{T1} \geq T_1 | P_{T0}\} &= \alpha_{T_1} \\ P\{Y_{T2} \geq T_2, Y_{T1} < T_1 | P_{T0}\} &= \alpha_{T_2} - \alpha_{T_1} \\ &\vdots \\ P\{Y_{TK} \geq T, Y_{T1} < T_1, Y_{T2} < T_2, \ldots, Y_{T(K-1)} < T_{K-1}, | P_{T0}\} &= \alpha_{T_K} - \alpha_{T_{K-1}} \end{aligned}$$

For each n, the toxicity boundaries are fixed with the usual group sequential theory. Then the boundaries and interim group sizes associated with response evaluation are searched for. The

process is repeated over a range of sample sizes. This is an extremely flexible design which can include the two-stage bivariate design similar to the Conaway and Petroni methodology, as well as the continuous toxicity monitoring.

We propose to search for the response boundaries and total sample size, n, given the pre-specified α, β, P_{R_0}, P_{R_a}, P_{T_0}, and P_{T_a} along with the toxicity boundaries based on α_T and ρ. In other words, for possible sample sizes we can fix the toxicity boundaries, then search for the response boundaries that maintain the specified α and β. The design with the desired characteristics is selected from the collection of possible designs. The design which minimizes the expected or the total sample size can be selected but designs with other characteristics can be selected as well.

4.3 Design Based on Data Monitoring Committee's Meeting Schedule

There are many possible designs to consider, such as a bivariate design that monitors response and toxicity in two stages at the same time. It is also possible to construct the continuous toxicity monitoring design, which monitors response in two stages. We will focus on a design that monitors toxicity four times and response twice, at the second and forth toxicity monitoring. The toxicity monitoring is equally spaced. In reality, this will correspond to a data monitoring committee's meeting schedule but the response is measured after several cycles of treatment in two stages. Since there are only four toxicity evaluations, the first response evaluation will be denoted as Y_{Rn_2} to indicate that it includes all n_2 subjects.

The critical region is

$$\begin{aligned} CR_{RR} = \{&(Y_{Tn_1},\ Y_{Tn_2},\ Y_{Tn_3},\ Y_{Tn};\ Y_{Rn_2},\ Y_{Rn}) : \\ &Y_{Tn_1} < T_1,\ Y_{Tn_2} < T_2,\ Y_{Tn_3} < T_3,\ Y_{Tn} < T, Y_{Rn_2} > R_2,\ Y_{Rn} > R\}. \end{aligned}$$

The constraints are

$$\begin{aligned} P\{(Y_{Tn_1}, Y_{Tn_2}, Y_{Tn_3}, Y_{Tn_4}; Y_{Rn_2}, Y_{Rn}) \in CR_{RR} | (P_{R_0}, P_{T_0}), \rho\} &\leq \alpha \\ P\{(Y_{Tn_1}, Y_{Tn_2}, Y_{Tn_3}, Y_{Tn_4}; Y_{Rn_2}, Y_{Rn}) \in CR_{RR} | (P_{R_A}, P_{T_A}), \rho\} &\geq 1 - \beta \\ \sum_{k=1}^{4} \alpha_{T_k} &\leq \alpha_T \end{aligned}$$

where T_1, T_2, T_3, and T are the toxicities boundaries based on the usual group sequential theory. The toxicity boundary values are found through a search algorithm of possible boundary values. The starting point is the discrete transformation of the boundary values created by the Pocock procedure. The Pocock procedure returns values $\{a_1,\ a_2,\ a_3,\ a_4\}$ that are associated with the standardized test statistics $\{Z_1,\ Z_2,\ Z_3,\ Z_4\}$. The transformed boundary values will be

denoted as $\{b_1,\ b_2,\ b_3,\ b_4\}$ but the alpha-level maybe larger than specified. The following formula was used to modify the standardized boundary values

$$b_k = \left\lceil \left(a_k \sqrt{\frac{P_{T_0}(1 - P_{T_0})}{k \cdot n}} + P_{T_0} \right) k \cdot n \right\rceil . \tag{15}$$

Then all possible values between 1 to $(b_k + 10)$ for $k = 1, 2, 3, 4$ were calculated and all combinations created with $b_1 \leq b_2 \leq b_3 \leq b_4$. The set of boundary values with the smallest squared difference from the specified alpha-level was selected and denoted as $\{T_1,\ T_2,\ T_3,\ T\}$.

Once the toxicity boundary values are determined, a search can be performed for the response boundaries. In a manner similar to Simon's 2-Stage design, we can find boundary values that either minimize the total sample size or the expected sample size. There are many designs available in Table 7 with different combinations of input parameters.

TABLE 7

Phase II Design That Monitors Toxicity Four Times and Response at the Second and Forth Time

Criteria				Optimal Design								Characteristics		
P_{T_0}	P_{T_A}	P_{R_0}	P_{R_a}	T_1	T_2	T_3	T	R_2	n_2	R	N	PET	ASN	Power
$\beta = 0.1$, $\alpha = 0.05$, $\alpha_T = 0.05$														
0.33	0.30	0.1	0.3	7	11	16	26	2	20	7	40	0.693	27.0	0.901
0.33	0.30	0.2	0.4	9	19	19	24	5	26	15	52	0.592	36.6	0.909
0.33	0.30	0.3	0.5	12	18	24	28	11	32	25	64	0.775	39.2	0.901
0.33	0.30	0.4	0.6	14	19	23	32	15	34	34	68	0.762	42.3	0.901
0.33	0.30	0.5	0.7	9	19	19	24	17	26	36	52	0.782	31.6	0.901
0.33	0.25	0.1	0.3	7	13	20	22	3	22	7	44	0.835	25.2	0.914
0.33	0.25	0.2	0.4	9	19	19	24	6	26	15	52	0.756	32.3	0.902
0.33	0.25	0.3	0.5	10	16	25	25	9	28	22	56	0.686	36.8	0.907
0.33	0.25	0.4	0.6	10	16	25	25	11	28	28	56	0.555	40.4	0.909
0.33	0.25	0.5	0.7	12	18	24	28	18	32	38	64	0.814	38.0	0.907
0.20	0.17	0.1	0.3	7	10	12	14	3	22	7	44	0.833	25.7	0.908
0.20	0.17	0.2	0.4	7	10	15	19	6	28	16	56	0.694	36.4	0.916
0.20	0.17	0.3	0.5	9	12	15	21	11	32	25	64	0.783	39.0	0.903
0.20	0.17	0.4	0.6	8	14	16	21	15	34	34	68	0.760	42.1	0.904
0.20	0.17	0.5	0.7	9	13	15	18	16	30	36	60	0.715	38.6	0.907
0.20	0.15	0.1	0.3	7	10	12	14	3	22	7	44	0.832	25.7	0.917
0.20	0.15	0.2	0.4	7	12	14	16	6	26	15	52	0.753	32.4	0.900
0.20	0.15	0.3	0.5	7	10	15	19	9	28	22	56	0.698	36.3	0.902
0.20	0.15	0.4	0.6	7	10	15	19	11	28	28	56	0.572	39.8	0.903
0.20	0.15	0.5	0.7	9	13	15	18	38	32	38	64	0.821	37.8	0.904
$\beta = 0.1$, $\alpha = 0.10$, $\alpha_T = 0.05$														
0.33	0.30	0.1	0.3	8	9	14	18	1	16	5	32	0.538	23.4	0.910
0.33	0.30	0.2	0.4	7	13	20	22	4	22	12	44	0.562	31.3	0.907
0.33	0.30	0.3	0.5	5	10	13	16	8	26	19	52	0.641	35.4	0.923
0.33	0.30	0.4	0.6	8	13	20	23	9	24	23	48	0.598	35.6	0.910
0.33	0.30	0.5	0.7	7	13	20	22	13	22	30	44	0.565	31.2	0.922
0.33	0.25	0.1	0.3	8	13	20	23	2	18	10	36	0.745	22.6	0.905
0.33	0.25	0.2	0.4	7	11	14	18	2	18	10	36	0.301	30.7	0.904
0.33	0.25	0.3	0.5	8	13	20	23	7	24	18	48	0.581	33.9	0.924
0.33	0.25	0.4	0.6	8	13	20	23	10	24	23	48	0.663	31.9	0.906
0.33	0.25	0.5	0.7	9	19	19	24	14	26	30	52	0.731	32.0	0.917

Criteria				Optimal Design								Characteristics		
P_{T_0}	P_{T_A}	P_{R_0}	P_{R_a}	T_1	T_2	T_3	T	R_2	n_2	R	N	PET	ASN	Power
$\beta = 0.1$, $\alpha = 0.10$, $\alpha_T = 0.05$														
0.20	0.17	0.1	0.3	5	8	10	11	1	16	5	32	0.526	23.5	0.918
0.20	0.17	0.2	0.4	8	11	11	13	3	20	11	40	0.467	31.6	0.909
0.20	0.17	0.3	0.5	7	10	15	15	7	24	18	48	0.571	34.2	0.915
0.20	0.17	0.4	0.6	7	12	14	16	11	26	25	52	0.680	34.3	0.903
0.20	0.17	0.5	0.7	7	12	14	16	14	26	30	52	0.727	33.0	0.907
0.20	0.15	0.1	0.3	8	8	11	12	2	18	6	36	0.739	22.7	0.904
0.20	0.15	0.2	0.4	8	8	11	12	2	18	10	36	0.287	30.9	0.902
0.20	0.15	0.3	0.5	7	10	15	15	7	24	18	48	0.571	34.2	0.924
0.20	0.15	0.4	0.6	7	10	15	15	10	24	23	48	0.657	32.2	0.905
0.20	0.15	0.5	0.7	7	12	14	16	14	26	30	52	0.727	33.0	0.915
$\beta = 0.2$, $\alpha = 0.05$, $\alpha_T = 0.05$														
0.33	0.30	0.1	0.3	8	9	14	18	2	16	6	32	0.800	19.2	0.816
0.33	0.30	0.2	0.4	7	13	20	22	6	22	12	44	0.873	24.4	0.805
0.33	0.30	0.3	0.5	7	13	20	22	7	22	18	44	0.685	28.5	0.810
0.33	0.30	0.4	0.6	8	13	20	23	11	24	24	48	0.795	28.7	0.817
0.33	0.30	0.5	0.7	7	11	14	18	11	18	26	36	0.641	24.5	0.803
0.33	0.25	0.1	0.3	8	9	14	18	2	16	6	32	0.800	19.2	0.832
0.33	0.25	0.2	0.4	7	11	14	18	4	18	11	36	0.728	22.8	0.800
0.33	0.25	0.3	0.5	8	13	20	23	9	24	19	48	0.853	27.3	0.806
0.33	0.25	0.4	0.6	8	13	20	23	11	24	24	48	0.795	28.7	0.832
0.33	0.25	0.5	0.7	7	13	20	23	12	22	27	44	0.749	27.1	0.818
0.20	0.17	0.1	0.3	5	8	10	11	2	16	6	32	0.794	19.2	0.823
0.20	0.17	0.2	0.4	7	10	12	14	6	22	12	44	0.870	24.9	0.814
0.20	0.17	0.3	0.5	7	10	12	14	7	22	18	44	0.679	29.1	0.820
0.20	0.17	0.4	0.6	7	10	15	15	11	24	24	48	0.790	29.0	0.824
0.20	0.17	0.5	0.7	7	10	12	14	12	22	27	44	0.745	27.6	0.813
0.20	0.15	0.1	0.3	5	8	10	11	2	16	6	32	0.794	19.2	0.831
0.20	0.15	0.2	0.4	8	11	11	13	5	20	12	40	0.809	23.7	0.806
0.20	0.15	0.3	0.5	7	10	15	15	9	24	19	48	0.850	27.6	0.806
0.20	0.15	0.4	0.6	7	10	15	15	11	24	24	48	0.790	29.0	0.832
0.20	0.15	0.5	0.7	7	10	15	15	14	24	29	48	0.849	27.6	0.805

Criteria				Optimal Design								Characteristics		
P_{T_0}	P_{T_A}	P_{R_0}	P_{R_a}	T_1	T_2	T_3	T	R_2	n_2	R	N	PET	ASN	Power
$\beta = 0.2$, $\alpha = 0.10$, $\alpha_T = 0.05$														
0.33	0.30	0.1	0.3	5	10	13	16	2	14	4	28	0.849	15.8	0.802
0.33	0.30	0.2	0.4	5	10	13	16	2	14	8	28	0.474	21.1	0.814
0.33	0.30	0.3	0.5	5	10	13	16	2	14	11	28	0.200	24.9	0.801
0.33	0.30	0.4	0.6	7	11	16	26	9	20	19	40	0.768	24.5	0.822
0.33	0.30	0.5	0.7	8	9	14	18	10	16	22	32	0.687	21.0	0.846
0.33	0.25	0.1	0.3	5	10	13	16	2	14	4	28	0.849	15.8	0.817
0.33	0.25	0.2	0.4	6	9	10	13	1	12	7	24	0.305	20.5	0.800
0.33	0.25	0.3	0.5	5	10	13	16	3	14	11	28	0.385	22.3	0.811
0.33	0.25	0.4	0.6	8	9	14	18	6	16	16	32	0.362	26.2	0.813
0.33	0.25	0.5	0.7	8	9	14	18	8	16	19	32	0.618	22.1	0.834
0.20	0.17	0.1	0.3	4	8	8	9	1	12	4	24	0.669	15.9	0.830
0.20	0.17	0.2	0.4	6	6	11	11	2	14	8	28	0.472	21.4	0.821
0.20	0.17	0.3	0.5	6	6	11	11	3	14	11	28	0.383	22.6	0.804
0.20	0.17	0.4	0.6	5	8	10	11	6	16	16	32	0.538	23.3	0.805
0.20	0.17	0.5	0.7	5	8	10	11	8	16	19	32	0.607	22.2	0.825
0.20	0.15	0.1	0.3	4	8	8	9	1	12	4	24	0.669	15.0	0.834
0.20	0.15	0.2	0.4	6	6	11	11	2	14	8	28	0.472	21.4	0.830
0.20	0.15	0.3	0.5	6	6	11	11	3	14	11	28	0.383	22.6	0.812
0.20	0.15	0.4	0.6	5	8	10	11	6	16	16	32	0.538	23.3	0.813
0.20	0.15	0.5	0.7	5	8	10	11	8	16	19	32	0.607	22.2	0.833

4.4 Simulations

A series of simulations was constructed to evaluate the operating characteristics of the bivariate test procedure when the correlation is different than specified. Designs were selected from Table 7 to be evaluated through the simulation. The simulations will also be used to examine the accuracy of the analytic expressions. The simulation is conducted in the following manner:

- Generate n observations from a correlated bivariate binomial distribution with marginal probabilities $P_R = P(X_{Rk} = 1)$ and $P_T = P(X_{Tk} = 1)$.
- Compare the cumulative number of toxicities observed through the first n_1 patients to the toxicity boundary value T_1. If the number of toxicities is equal to or exceeds the boundary value, then stop the trial; otherwise, enroll the next $(n_2 - n_1)$ patients.
- If the trial did not terminate, then at observation n_2 sum up the total number of responses. If there are R_2 or fewer responses, then we halt the trial. Also, compare the number of toxicities through patient n_2 to the associated boundary T_2. If Y_{Tn_2} is equal to or exceeds the boundary, then discontinue the study.

- If the trial did not stop, then enroll the additional $(n_3 - n_2)$ subjects. If the cumulative number of toxicities through patient n_3 is greater than $T_3 - 1$, then stop the trial; otherwise, enroll the remaining $(n - n_3)$ patients.
- Finally, after all n subjects are enrolled we do not reject the null hypothesis if there are R or few responses or if the number of toxicities exceeds $T - 1$.

The above procedure is repeated 100,000 times for each value $m \cdot P(X_{Rk} = 1 \text{ and } X_{Tk} = 1) = m \cdot P_{11}$ where $m \in \{0, 0.01, 0.02, \ldots, 0.98, 0.99, 1.00\}$. The sequence of joint probabilities produces the smooth figures to be discussed. The joint probability is transformed into the correlation through the following relationship:

$$\rho(X_{Rk}, X_{Tk}) = \frac{P_{11} - P_R P_T}{\sqrt{P_R(1 - P_R)P_T(1 - P_T)}}$$

The range of the correlation between response and toxicity is not [-1, 1] but is limited by joint probability P_{11}. When $P_{11} = 0$ the correlation will be the smallest at $\frac{-P_R P_T}{\sqrt{P_R(1-P_R)P_T(1-P_T)}}$. The largest value for $P_{11} = min(P_R,\ P_T) > P_R \cdot P_T$ and results in the largest value for $\rho(X_{Rk}, X_{Tk})$.

The power analysis is performed in a two step process. The first step is to determine all combinations of probabilities associated with response and toxicity. This is completed by creating two sequences from 0 to 1 by 0.01 increments. Then all possible combinations of the two sequences are determined, which allows the complete power surface to be specified. The power surface is determined under the assumption that toxicity and response are independent. Finally, the effect of the correlation on the power is analyzed.

4.5 Simulation Results

The simulation results reported are based on the first row of Table 7. The nominal size of the bivariate procedure is pictured in Figure 9. The design minimizes the expected sample size and was created under the assumption of independence between toxicity and response. The assumption is equivalent to a correlation of 0. We can see that that bivariate design becomes more conservative as the correlation increases from 0 to the maximum of 93.3 %. It is also important to note that the simulated type I error rate increases as the correlation goes from 0 to the minimum value but it does not exceed the pre-specified type I error rate.

The average sample size (ASN) is pictured in Figure 10. The ASN is the largest when the correlation is the smallest and decreases as the correlation increases. The largest expected sample size is 27.1 and decreases to approximately 26.6 when the correlation between response and toxicity is at the maximum value.

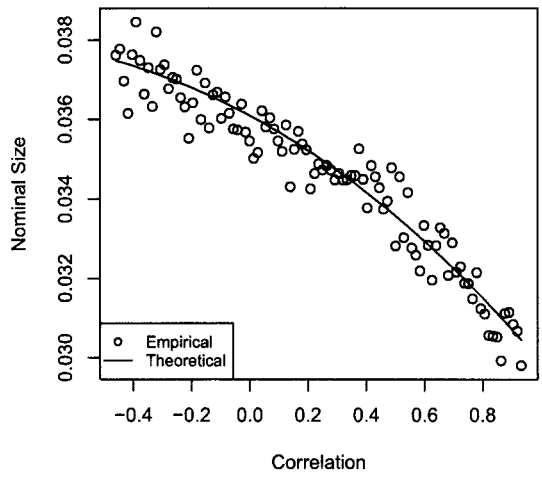

Figure 9. Effect of Correlation on the Type I Error of the Bivariate Test with $P_{R_0} = 0.1$, $P_{R_A} = 0.3$, $P_{T_0} = 0.33$, $P_{T_A} = 0.30$, $\alpha = 0.05$, $\alpha_T = 0.05$ and $\beta = 0.10$

Figure 11 graphically displays the effect of the correlation on the probability of early termination (PET). The PET is the smallest when the correlation is the smallest and increases as the correlation increases. The PET reaches its maximum value of 71 % when the correlation obtains its maximum value of 93.3%.

The power of the bivariate design is pictured in Figure 12. The bivariate procedure successfully identifies treatments that induce a large response rate and low toxicity rate. We can also see that the combined bivariate procedure's nominal size is less than 5% when the response rate is less than or equal to 10 % and the toxicity rate is larger than 33 %.

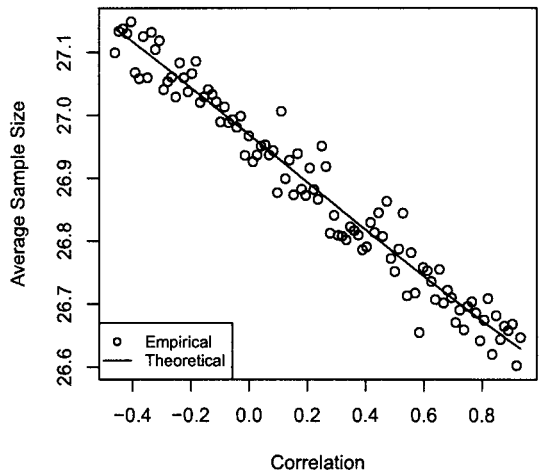

Figure 10. Effect of Correlation on the Average Sample Size of the Bivariate Test with $P_{R_0} = 0.1$, $P_{R_A} = 0.3$, $P_{T_0} = 0.33$, $P_{T_A} = 0.30$, $\alpha = 0.05$, $\alpha_T = 0.05$ and $\beta = 0.10$

The effect of the correlation on the bivariate procedure's power function is displayed in Figure 13. The power of the bivariate test is affected by misspecification of the correlation but we can see it never goes below the pre-specified level in this example. The trial is designed under the assumption of independence with an associated power of approximately 90.1 %. The power decreases to approximately 90% when the correlation achieves the maximum values.

The ability to control the type I error rate over the entire null hypothesis region is difficult with the typical sample sizes expected in phase II single-arm trials. The design is able to control the type I and II error rates when both endpoints are either in the null hypothesis region or in the alternative hypothesis region. Next, we examine the type I error rate when one endpoint is in the rejection region while the other endpoint is in the acceptance region. There are two possible cases.

First, we will consider a toxicity rate that is smaller than expected under the null hypothesis and allow the response rate to vary the over range of possible response values. In other words, the treatment does not produce enough toxic events to declare it unsafe and we consider many possible response rates. Figure 14 displays the power curve when the toxicity rate is fixed at $P_T = 0.25$ and the response rate is allowed to vary between 5% and 70%. The parameters used to construct the specific design are $P_{R_0} = 0.1$ and $P_{R_A} = 0.3$. We can see that the design achieves the nominal size specified when the response rate is 10%. The power reaches 90% when the response rate attains 30%. This is the expected result based on the design parameters.

The next scenario to consider is when the null response rate is larger than expected under

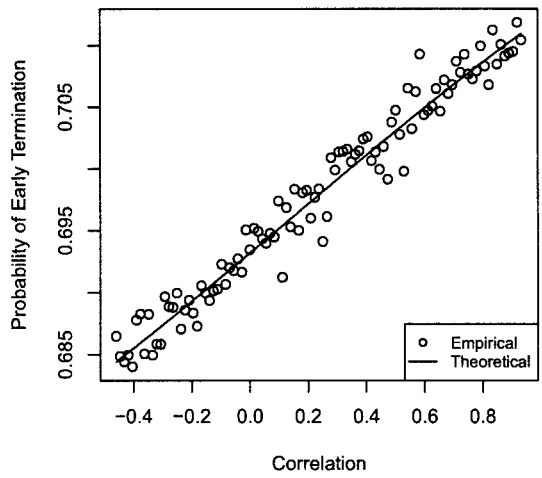

Figure 11. Effect of Correlation on the Probability of Early Termination of the Bivariate Test with $P_{R_0} = 0.1$, $P_{R_A} = 0.3$, $P_{T_0} = 0.33$, $P_{T_A} = 0.30$, $\alpha = 0.05$, $\alpha_T = 0.05$ and $\beta = 0.10$

the null hypothesis and the toxicity rates vary over the set of possible values. The response rate is set at 30% and the power curve is over all possible toxicity rates between 5% and 70%. The curve is depicted in Figure 15. We can see that the curve decreases towards the nominal size as the toxicity rate increases beyond 33% but not as fast as desired. It is possible to declare a treatment effective with a larger than expected toxicity rate. This is not necessarily a bad thing and the consequences will be discussed in the Section 4.6.

Now, we will compare the traditional Simon 2-Stage design to the total sample size and expected sample size from the proposed bivariate design. Table 8 contains the results of the Minimax and Optimal designs along with expected sample from the bivariate design. The greatest increase in the sample size between the Simon 2-Stage design and the new bivariate design is at most 35.7% for the set of parameters considered. The largest increase, due to the additional endpoint, is associated with the smaller sample sizes created when power is set at 80% and the type I error rate is 10 %. The actual increase in the expected sample size is at the most 6.3 subjects while the actual increase in the maximum number of patients is 10.

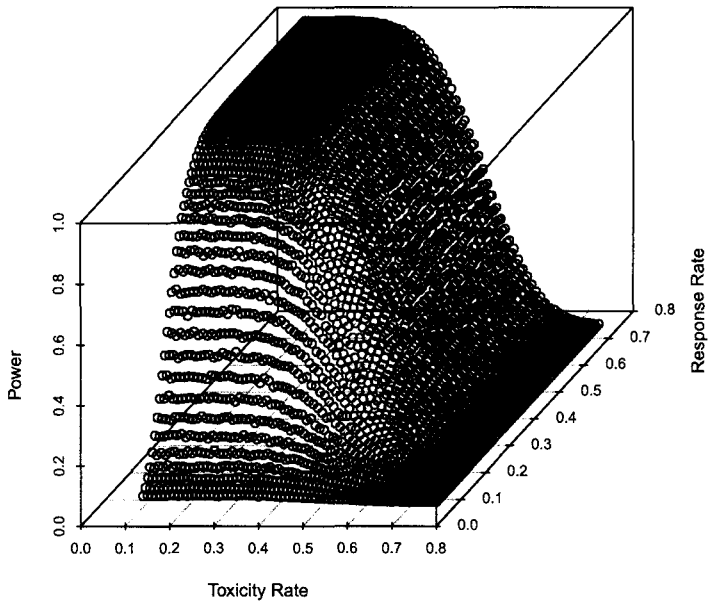

Figure 12. Power Surface of the Bivariate Test with $P_{R_0} = 0.1$, $P_{R_A} = 0.3$, $P_{T_0} = 0.33$, $P_{T_A} = 0.30$, $\alpha = 0.05$, $\alpha_T = 0.05$ and $\beta = 0.10$

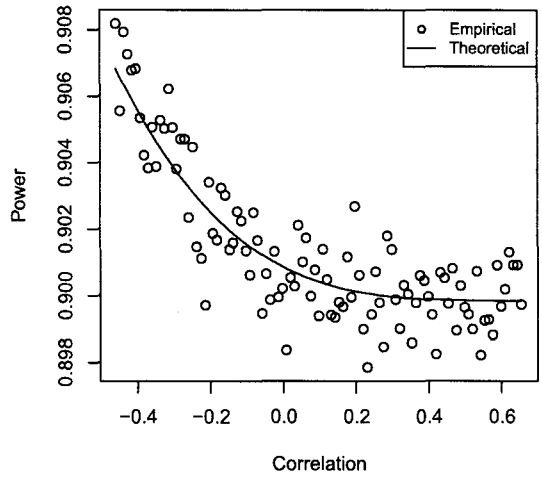

Figure 13. Effect of Correlation on the Power of the Bivariate Test with $P_{R_0} = 0.1$, $P_{R_A} = 0.3$, $P_{T_0} = 0.33$, $P_{T_A} = 0.30$, $\alpha = 0.05$, $\alpha_T = 0.05$ and $\beta = 0.10$

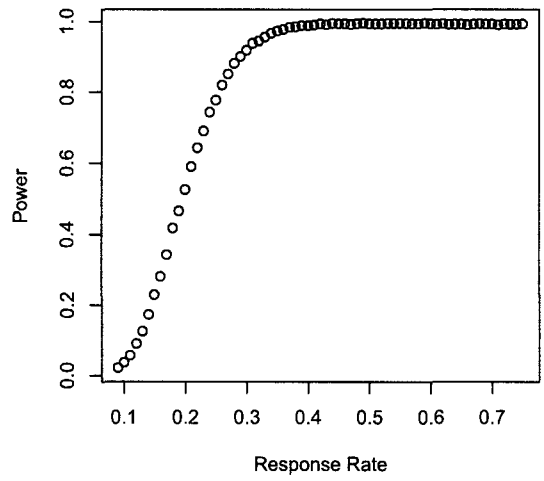

Figure 14. Marginal Power Curve Over the Response Rates when Toxicity is Fixed in the Alternative $P_{R_0} = 0.1$, $P_{R_A} = 0.3$, $P_{T_0} = 0.33$, $P_{T_A} = 0.30$, $\alpha = 0.05$, $\alpha_T = 0.05$ and $\beta = 0.10$

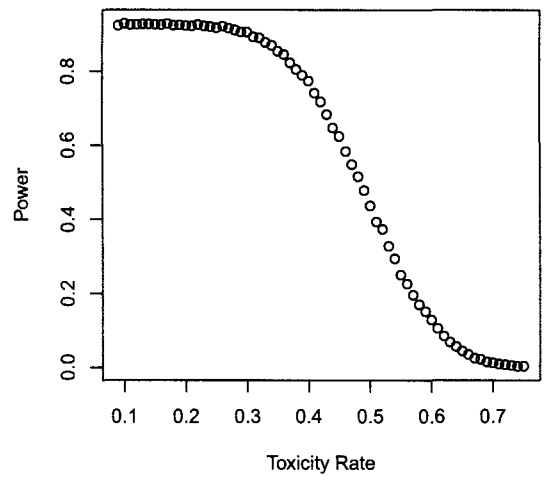

Figure 15. Marginal Power Curve Over the Toxicity Rates when Response is Fixed in the Alternative $P_{R_0} = 0.1$, $P_{R_A} = 0.3$, $P_{T_0} = 0.33$, $P_{T_A} = 0.30$, $\alpha = 0.05$, $\alpha_T = 0.05$ and $\beta = 0.10$

TABLE 8

Simon 2-Stage Optimal, Minimax, and Bivariate Design Total Samples and Average Sample Sizes (ASN)

Criteria				Optimal		Minimax		Bivariate Design	
P_{T_0}	P_{T_A}	P_{R_0}	P_{R_a}	N	ASN	N	ASN	N	ASN
$\beta = 0.1$, $\alpha = 0.05$, $\alpha_T = 0.05$									
0.33	0.30	0.1	0.3	35	22.5	33	26.2	40	27.0
0.33	0.25	0.1	0.3	35	22.5	33	26.2	44	25.2
0.20	0.17	0.1	0.3	35	22.5	33	26.2	44	25.7
0.20	0.15	0.1	0.3	35	22.5	33	26.2	44	25.7
0.33	0.30	0.2	0.4	54	30.4	45	31.2	52	36.6
0.33	0.25	0.2	0.4	54	30.4	45	31.2	52	32.3
0.20	0.17	0.2	0.4	54	30.4	45	31.2	56	36.4
0.20	0.15	0.2	0.4	54	30.4	45	31.2	52	32.4
0.33	0.30	0.3	0.5	63	34.7	53	36.6	64	39.2
0.33	0.25	0.3	0.5	63	34.7	53	36.6	56	36.8
0.20	0.17	0.3	0.5	63	34.7	53	36.6	64	39.0
0.20	0.15	0.3	0.5	63	34.7	53	36.6	56	36.3
0.33	0.30	0.4	0.6	66	36.0	54	38.1	68	42.3
0.33	0.25	0.4	0.6	66	36.0	54	38.1	56	40.4
0.20	0.17	0.4	0.6	66	36.0	54	38.1	68	42.1
0.20	0.15	0.4	0.6	66	36.0	54	38.1	56	39.8
0.33	0.30	0.5	0.7	61	34.0	53	36.1	52	31.6
0.33	0.25	0.5	0.7	61	34.0	53	36.1	64	38.0
0.20	0.17	0.5	0.7	61	34.0	53	36.1	60	38.6
0.20	0.15	0.5	0.7	61	34.0	53	36.1	64	37.8

Criteria				Optimal		Minimax		Bivariate Design	
P_{T_0}	P_{T_A}	P_{R_0}	P_{R_a}	N	ASN	N	ASN	N	ASN
$\beta = 0.1$, $\alpha = 0.10$, $\alpha_T = 0.05$									
0.33	0.30	0.1	0.3	35	19.8	25	20.4	32	23.4
0.33	0.25	0.1	0.3	35	19.8	25	20.4	36	22.6
0.20	0.17	0.1	0.3	35	19.8	25	20.4	32	23.5
0.20	0.15	0.1	0.3	35	19.8	25	20.4	36	22.7
0.33	0.30	0.2	0.4	37	26.0	36	28.3	44	31.3
0.33	0.25	0.2	0.4	37	26.0	36	28.3	36	30.7
0.20	0.17	0.2	0.4	37	26.0	36	28.3	40	31.6
0.20	0.15	0.2	0.4	37	26.0	36	28.3	36	30.9
0.33	0.30	0.3	0.5	46	29.9	39	35.0	52	35.4
0.33	0.25	0.3	0.5	46	29.9	39	35.0	48	33.9
0.20	0.17	0.3	0.5	46	29.9	39	35.0	48	34.2
0.20	0.15	0.3	0.5	46	29.9	39	35.0	48	34.2
0.33	0.30	0.4	0.6	46	30.2	41	33.8	48	35.6
0.33	0.25	0.4	0.6	46	30.2	41	33.8	48	31.9
0.20	0.17	0.4	0.6	46	30.2	41	33.8	52	34.3
0.20	0.15	0.4	0.6	46	30.2	41	33.8	48	32.2
0.33	0.30	0.5	0.7	45	29.0	39	31.0	44	31.2
0.33	0.25	0.5	0.7	45	29.0	39	31.0	52	32.0
0.20	0.17	0.5	0.7	45	29.0	39	31.0	52	33.0
0.20	0.15	0.5	0.7	45	29.0	39	31.0	52	33.0

Criteria				Optimal		Minimax		Bivariate Design	
P_{T_0}	P_{T_A}	P_{R_0}	P_{R_a}	N	ASN	N	ASN	N	ASN
0.33	0.30	0.1	0.3	29	15.0	25	19.5	32	19.2
0.33	0.25	0.1	0.3	29	15.0	25	19.5	32	19.2
0.20	0.17	0.1	0.3	29	15.0	25	19.5	32	19.2
0.20	0.15	0.1	0.3	29	15.0	25	19.5	32	19.2
0.33	0.30	0.2	0.4	43	20.6	33	22.3	44	24.4
0.33	0.25	0.2	0.4	43	20.6	33	22.3	36	22.8
0.20	0.17	0.2	0.4	43	20.6	33	22.3	44	24.9
0.20	0.15	0.2	0.4	43	20.6	33	22.3	40	23.7
0.33	0.30	0.3	0.5	46	23.6	39	25.7	44	28.5
0.33	0.25	0.3	0.5	46	23.6	39	25.7	48	27.3
0.20	0.17	0.3	0.5	46	23.6	39	25.7	44	29.1
0.20	0.15	0.3	0.5	46	23.6	39	25.7	48	27.6
0.33	0.30	0.4	0.6	46	24.5	39	34.4	48	28.7
0.33	0.25	0.4	0.6	46	24.5	39	34.4	48	28.7
0.20	0.17	0.4	0.6	46	24.5	39	34.4	48	29.0
0.20	0.15	0.4	0.6	46	24.5	39	34.4	48	29.0
0.33	0.30	0.5	0.7	43	23.5	37	27.7	36	24.5
0.33	0.25	0.5	0.7	43	23.5	37	27.7	44	27.1
0.20	0.17	0.5	0.7	43	23.5	37	27.7	44	27.6
0.20	0.15	0.5	0.7	43	23.5	37	27.7	48	27.6

Criteria				Optimal		Minimax		Bivariate Design	
P_{T_0}	P_{T_A}	P_{R_0}	P_{R_a}	N	ASN	N	ASN	N	ASN
$\beta = 0.2$, $\alpha = 0.10$, $\alpha_T = 0.05$									
0.33	0.30	0.1	0.3	18	12.7	18	12.7	28	15.8
0.33	0.25	0.1	0.3	18	12.7	18	12.7	28	15.8
0.20	0.17	0.1	0.3	18	12.7	18	12.7	24	15.9
0.20	0.15	0.1	0.3	18	12.7	18	12.7	24	15.0
0.33	0.30	0.2	0.4	25	17.7	24	19.5	28	21.1
0.33	0.25	0.2	0.4	25	17.7	24	19.5	24	20.5
0.20	0.17	0.2	0.4	25	17.7	24	19.5	28	21.4
0.20	0.15	0.2	0.4	25	17.7	24	19.5	28	21.4
0.33	0.30	0.3	0.5	32	19.7	28	20.1	28	24.9
0.33	0.25	0.3	0.5	32	19.7	28	20.1	28	22.3
0.20	0.17	0.3	0.5	32	19.7	28	20.1	28	22.6
0.20	0.15	0.3	0.5	32	19.7	28	20.1	28	22.6
0.33	0.30	0.4	0.6	38	20.7	28	21.7	40	24.5
0.33	0.25	0.4	0.6	38	20.7	28	21.7	32	26.2
0.20	0.17	0.4	0.6	38	20.7	28	21.7	32	23.3
0.20	0.15	0.4	0.6	38	20.7	28	21.7	32	23.3
0.33	0.30	0.5	0.7	32	19.7	28	21.5	32	21.0
0.33	0.25	0.5	0.7	32	19.7	28	21.5	32	22.1
0.20	0.17	0.5	0.7	32	19.7	28	21.5	32	22.2
0.20	0.15	0.5	0.7	32	19.7	28	21.5	32	22.2

4.6 Discussion

The design sacrifices the ability to detect toxicity rates larger than desired when the response rate is also large enough to reject the null hypothesis. The trial will successfully identify agents with ineffective response with small toxicity. The data monitoring committee must carefully weight the cost of the additional response in terms of toxicity. Regardless of the results of the test, an agent that is efficacious maybe allowed to have a larger toxicity rate to achieve the result. In this situation, it is imperative that the data monitoring committee understand the consequences of the design, including the relatively small sample sizes, and make appropriate recommendations that consider more than just the statistical hypothesis test.

In the end, a trial that is very similar to the Simon 2-Stage design is possible that allows guidance for the data monitoring committee at each meeting. The additional cost of the guidance is small while the characteristics are preserved regardless of the toxicity rate. In large multiple center trials, the additional collection and analysis of the data maybe too burdensome, but the design can be modified to check both response and toxicity in two stages. The design also allows one to set toxicity as the priority instead of response, if the principal investigator is concerned with accurate monitoring of toxic events. The design can also incorporate other assumptions related to the correlation. The current algorithm requires a computing cluster with 25 nodes to search for the sample sizes. The next problems to consider include more efficient search algorithms and point estimation for response given the multistage, bivariate design. The flexible design also allows one to consider different monitoring schedules for response and toxicity.

CHAPTER 5

PROPER INFERENCE AFTER SINGLE-ARM, MULTISTAGE, BIVARIATE CLINICAL TRIALS

Treatments progress through phases in the drug evaluation process. The first clinical experiments, or phase I trials, are small studies intended to find the maximum tolerated dose (MTD). The MTD is typically the largest dose that is safe to administer. In the oncology setting, larger doses of cytotoxic agents usually result in larger response rates, but also increase the number of toxic events. The MTD will proceed into a phase II clinical trial designed to test the efficacy of the new agent against the standard treatment. The phase II trial can take on many different forms, but the most popular is the single-arm design (Stallard et al., 2001). If the new treatment is both safe and efficacious, then it will proceed into the larger phase III clinical trial. The phase III trial is a confirmatory study, whereas the phase II clinical trial is exploratory.

There are several single-arm, phase II clinical trial designs that incorporate multiple examinations of the data, as well as multiple endpoints. Conaway and Petroni (1995), as well as Bryant and Day (1995), both provide two-stage frequentist based designs that incorporate response and toxicity simultaneously. Ivanova et al. (2005) propose a phase II single-arm clinical trial design that measures response once at the end of the trial, but evaluates toxicity after each patient is enrolled. This is also referred to as continuous toxicity monitoring. Ray and Rai (2011b) expand the concept into a formal design that allows the toxicity to be monitored on a different schedule than response. The design also encompasses the continuous toxicity monitoring concept with two or more examinations of response. The design framework provides statistical guidance associated with toxicity monitoring. Further, the design allows multiple evaluations of the response rate during the conduct of the trial. Ray and Rai provide an example design that monitors toxicity four times. The design also evaluates response at the second and final toxicity examinations.

The design of the phase III clinical trial requires an estimate of the response rate based on the phase II trial results. Jung and Kim (2004) provide a uniformly minimum variance unbiased estimate (UMVUE) of the response rate for single-arm, multiple-stage trials that only examine response. They note that the usual maximum likelihood estimate (MLE) is biased due to the

optimal sampling effect. The bias is a consequence of the trial design, since it results in only observing extreme values due to crossing the pre-specified boundaries.

In this chapter, we investigate an estimator for response based on the clinical trial designs that include both response and toxicity. We also consider the ability to monitor the toxicity on a different schedule from response, which includes the continuous toxicity monitoring example. Section 5.1 contains the theoretical derivation of the estimator, which is explored through simulation in Section 5.3.

5.1 Distribution Theory

Analysis of the distribution of response and toxicities is required before we can determine expressions for an unbiased point estimator. Suppose that both response and toxic events are binary outcomes. Let $k \in \{1, 2, \ldots K\}$ be the number of stages in which the data will be evaluated. Let $n_1, \ldots n_K$ be the number of subjects accrued at each stage in the trial such that $n_1 + \cdots + n_K = n$ where n is total sample size. Let X_{rtk} be the number of observations associated with the k^{th} stage where r indicates if a response occurred and t indicates if a toxic event occurred. Then $X_{rtk} = (X_{00k},\ X_{01k},\ X_{10k},\ X_{11k})$ follows a multinomial distribution with parameters $P_{00},\ P_{01},\ P_{10}$, and P_{11}. The probability of a response is $P_R = P_{10} + P_{11}$ and the probability of a toxicity is $P_T = P_{01} + P_{11}$. Let $Y_{rtm} = \sum_{k=1}^{m} X_{rtk}$ be the accumulated data through the m^{th} stage for $r = 0, 1$ and $t = 0, 1$. The total number of responses through stage m will be defined as $Y_{Rm} = \sum_{k=1}^{m} \sum_{t=0}^{1} X_{1tk} = \sum_{k=1}^{m} X_{Rk}$ and the total number of toxicities through stage m will be $Y_{Tm} = \sum_{k=1}^{m} \sum_{r=0}^{1} X_{r1k} = \sum_{k=1}^{m} X_{Tk}$. The response boundary values will be denoted as R_k and the toxicity boundary values will be denoted as T_k. The trial should stop if $Y_{Rk} \leq R_k$ or if $Y_{Tk} \geq T_k$, for $k = 1, 2, ..., K$.

The full joint probability of Y_{Rm} responses and Y_{Tm} toxicities at stage m is

$$
\begin{aligned}
& f(m, y_{Rm}, y_{Tm}|\mathbf{P}) \\
& = Pr\{M = m, Y_{Rm} = y_{Rm}, Y_{Tm} = y_{Tm}|\mathbf{P}\} \\
& = Pr\{Y_{Rm} = y_{Rm}, Y_{Tm} = y_{Tm}, y_{Rk} \geq R_k + 1, y_{Tk} \leq T_k - 1, k = 1 \cdots (m-1)|\mathbf{P}\} \\
& = \sum_{R(m, y_{Rm}, y_{Tm})} Pr\{X_{R1} = x_{R1}, X_{T1} = x_{T1}, \cdots, X_{Rm} = x_{Rm}, X_{Tm} = x_{Tm}|\mathbf{P}\}
\end{aligned}
$$

where

$$
\begin{aligned}
R(m, y_{Rm}, y_{Tm}) \ = \{(x_{R1}, \cdots, x_{Rm};\ & x_{T1}, ..., x_{Tm}) : \\
& x_{R1} + \cdots + x_{Rm} = y_{Rm}, x_{T1} + \cdots + x_{Tm} = y_{Tm}, \\
& y_{Rk} \geq R_k + 1, y_{Tk} \leq T_k - 1, k = 1 \cdots (m-1)\}.
\end{aligned}
$$

The probability of Y_{Rm} responses at stage m is

$$f(m, y_{Rm}|\mathbf{P}) = \sum_{y_{Tm}} f(m, y_{Rm}, y_{Tm}|\mathbf{P})$$

so we will write

$$\begin{aligned} & f(m, y_{Rm}|\mathbf{P}) \\ & = Pr\{M = m, Y_{Rm} = y_{Rm}|\mathbf{P}\} \\ & = Pr\{Y_{Rm} = y_{Rm}, y_{Rk} \geq R_k + 1, y_{Tk} \leq T_k - 1, k = 1 \cdots (m-1)|\mathbf{P}\} \\ & = \sum_{R(m, y_{Rm})} Pr\{X_{R1} = x_{R1}, \cdots, X_{Rm} = x_{Rm}|\mathbf{P}\} \end{aligned} \tag{16}$$

where

$$\begin{aligned} R(m, y_{Rm}) \ = \{(x_{R1}, \cdots, x_{Rm}) : & x_{R1} + \cdots + x_{Rm} = y_{Rm}, \\ & y_{Rk} \geq R_k + 1, y_{Tk} \geq T_k - 1, k = 1 \cdots (m-1)\}. \end{aligned}$$

Let us examine

$$\begin{aligned} & Pr\{X_{R1} = x_{R1}, \cdots, X_{Rm} = x_{Rm}|\mathbf{P}\} = \\ & Pr\{X_{R1} = x_{R1}|\mathbf{P}\} \cdots Pr\{X_{Rm} = x_{Rm}|\mathbf{P}\} \end{aligned}$$

since the observations in each stage are independent.

Specifically, we will consider $Pr\{X_{Rm} = x_{Rm}|\mathbf{P}\}$ from any one stage.

$$P_R = P_{11} + P_{10} \Rightarrow$$

$$Pr\{X_{Rm} = x_{Rm}|\mathbf{P}\} = \sum_{B_{x_{Rm}}} f(x_{00m}, x_{01m}, x_{Rm}|\mathbf{P})$$

where

$$B_{x_{Rm}} = \{(x_{00m}, x_{01m}, x_{Rm}) : x_{00m} + x_{01m} = n_m - x_{Rm}\}.$$

$$
\begin{aligned}
& Pr\{X_{Rm} = x_{Rm} | \ \mathbf{P} \ \} \\
&= \sum_{B_{x_{Rm}}} n_m! \frac{P_R^{x_{Rm}}}{x_{Rm!}} \frac{P_{00}^{x_{00m}}}{x_{00m!}} \frac{P_{01m}^{x_{01m}}}{x_{01m!}} \\
&= \sum_{B_{x_{Rm}}} n_m! \frac{P_R^{x_{Rm}}}{x_{Rm!}} \frac{P_{00}^{x_{00m}}}{x_{00m!}} \frac{P_{01}^{x_{01m}}}{x_{01m!}} \frac{(n_m - x_{Rm})!}{(n_m - x_{Rm})!} \frac{(1-P_R)^{(n_m - x_{Rm})}}{(1-P_R)^{(n_m - x_{Rm})}} \\
&= \frac{n_i!}{(n_m - x_{Rm})! x_{Rm!}} P_R^{x_{Rm}} (1-P_R)^{(n_m - x_{Rm})} \\
&\qquad \sum_{B_{x_{Rm}}} \frac{(n_m - x_{Rm})!}{x_{00m}! x_{01m}!} \frac{P_{00}^{x_{00m}} P_{01}^{x_{01m}}}{(1-P_R)^{(n_m - x_{Rm})}} \\
&= \frac{n_m!}{(n_m - x_{Rm})! x_{Rm!}} P_R^{x_{Rm}} (1-P_R)^{(n_m - x_{Rm})} \\
&\qquad \sum_{B_{x_{Rm}}} \frac{(n_m - x_{Rm})!}{x_{00m}! x_{01m}!} \left\{ \frac{P_{00}}{(1-P_R)} \right\}^{x_{00m}} \left\{ \frac{P_{01}}{(1-P_R)} \right\}^{x_{01m}} \\
&= \frac{n_m!}{(n_m - x_{Rm})! x_{Rm!}} P_R^{x_{Rm}} (1-P_R)^{(n_m - x_{Rm})} \qquad (17)
\end{aligned}
$$

The summation is 1 because we are summing over all possible values of a multinomial distribution with parameters $n_m - x_{Rm}$, $\frac{P_{01}}{(1-P_R)}$, and $\frac{P_{00}}{(1-P_R)}$. This shows that the marginal distribution of a multinomial is a binomial distribution, which is required to simplify the calculations in the next step.

If we plug the result from Equation 17 into Equation 16, then

$$
\begin{aligned}
& f(m, y_{Rm} | \mathbf{P}) \\
&= \sum_{R(m,y_{Rm})} Pr\{X_{R1} = x_{R1}, \cdots, X_{Rm} = x_{Rm}\} \\
&= \sum_{R(m,y_{Rm})} \frac{n_1!}{(n_1 - x_{R1})! x_{R1!}} P_R^{x_{R1}} (1-P_R)^{(n_1 - x_{R1})} \cdots \\
&\qquad \frac{n_m!}{(n_m - x_{Rm})! x_{Rm!}} P_R^{x_{Rm}} (1-P_R)^{(n_m - x_{Rm})} \\
&= \sum_{R(m,y_{Rm})} \binom{n_1}{x_{R1}} \cdots \binom{n_m}{x_{Rm}} P_R^{\sum_{k=1}^m x_{Rk}} (1-P_R)^{\sum_{k=1}^m (n_k - x_{Rk})} \\
&= \sum_{R(m,y_{Rm})} \binom{n_1}{x_{R1}} \cdots \binom{n_m}{x_{Rm}} P_R^{y_{Rm}} (1-P_R)^{(\sum_{k=1}^m (n_k) - y_{Rm})}.
\end{aligned}
$$

This is a similar situation to that described by Jung and Kim (2004). Thus it seems reasonable to evaluate the resulting estimator with the additional restriction caused by the toxicity monitoring. The specific estimator under consideration is

$$
\tilde{P_R} = \frac{\sum \cdots \sum_{R(m,y_{Rm})} \binom{n_1 - 1}{x_{R1-1}} \binom{n_2}{x_{R2}} \cdots \binom{n_m}{x_{Rm}}}{\sum \cdots \sum_{R(m,y_{Rm})} \binom{n_1}{x_{R1}} \cdots \binom{n_m}{x_{Rm}}}
$$

where the summations in the numerator and the denominator are over

$$R(m, y_{Rm}) = \{(x_{R1}, \cdots, x_{Rm}) : x_{R1} + \cdots + x_{Rm} = y_{Rm}, \\ y_{Rk} \geq R_k + 1, y_{Tk} \leq T_k - 1, k = 1 \cdots (m-1)\}.$$

In other words, the trial can be stopped early for either toxicity or response considerations which is reflected in the additional restrictions.

The derived point estimator cannot be applied when the Simon 2-Stage design is combined with the continuous toxicity monitoring. The stage-wise sample sizes are one and the response is either a one or a zero. In either case, $\binom{1}{0} = \binom{1}{1} = 1$ and the proposed point estimator is not applicable. Therefore, we propose to modify the UMVUE by ignoring the toxicity monitoring or just applying the estimator proposed by Jung and Kim (2004). The specific estimator is

$$\widetilde{\widetilde{P}_R} = \frac{\sum_{\widetilde{R}(m,y_{Rm})} \binom{n_1-1}{x_{R1}-1}\binom{n_2}{x_{R2}} \cdots \binom{n_m}{x_{Rm}}}{\sum_{\widetilde{R}(m,y_{Rm})} \binom{n_1}{x_{R1}} \cdots \binom{n_m}{x_{Rm}}}$$

where

$$\widetilde{R}(m, y_{Rm}) = \{(x_{R1}, \cdots, x_{Rm}) : x_{R1} + \cdots + x_{Rm} = y_{Rm}, \\ y_{Rk} \geq R_k + 1, k = 1 \cdots (m-1)\}.$$

5.2 UMVUE Proof

The proof that this is the UMUVE starts with the probability mass function of m and Y_{Rm} which is

$$f(m, y_{Rm}|\mathbf{P}) = \sum_{R(m,y_{Rm})} \binom{n_1}{x_{R1}} \cdots \binom{n_m}{x_{Rm}} P_R^{y_{Rm}} (1 - P_R)^{(\sum_{k=1}^{m}(n_k) - y_{Rm})}.$$

We will write as

$$f(m, y_{Rm}|\mathbf{P}) = C_{(m,y_{Rm},y_{Tm})} P_R^{y_{Rm}} (1 - P_R)^{(\sum_{k=1}^{m}(n_k) - y_{Rm})}$$

with support

$$\omega = \bigcup_{m=1}^{K} \omega_m$$

where

$$\omega_m = \{(m, y_{Rm}) : (R_{m-1} + 1 \leq y_{Rm} \leq R_m) \cap y_{Tm} \leq T_{m-1} - 1\}.$$

The statistic (m, y_{Rm}) is sufficient by the factorization theorem.

Now, we prove the completeness of (m, y_{Rm}). In a manner similar that used by Jung and Kim, consider $h(P_R) = E_{P_R}\{g(m, y_{Rm})\}$ which is obtained as

$$\sum_{m=1}^{K} \sum_{y_{Rm}=r_{m-1}+1}^{r_m} g(m, y_{Rm}) f(m, y_{Rm}|P_R)$$

$$= \sum_{m=1}^{K} \sum_{y_{Rm}=r_{m-1}+1}^{r_m} g(m, y_{Rm}) C_{(m,y_{Rm},y_{Tm})} P_R^{y_{Rm}} (1-P_R)^{(\sum(n_i)-y_{Rm})} \tag{18}$$

Now, we need to show that $h(P_R) = 0$ for $p \in [0,1] \Rightarrow g(m, y_R) = 0$ for all $(m, y_R) \in (M, Y_R)$. Define 0^0 to be 1. If $P_R = 0$, then, from Equation 18 we have $g(1,0) = 0$. If $1 - P_R = 0$, then, from Equation 18 we have $g(1,n) = 0$. Now, for $P_R \in (0,1)$ let $P_j(P_R) = h(P_R)/P_R^j$ and $Q_l(P_R) = h(P_R)/(1-P_R)^l$. Each term, say term i, in Equation 18 has the factor $P_R^{j_i}(1-P_R)^{l_i}$ for some nonnegative integers j_i and l_i. Since all the terms have different factors $(j_i, l_i) \neq (j_j, l_j)$ if $i \neq j$, any subset of the terms in Equation 18 has a unique minimum either among the j_i's or the l_i's. If j_i's have a unique minimum j, then since $P_j(P_R) = 0$ for all $p \in (0,1)$, letting $p \to 0$ shows $g(m, y_R) = 0$ where $g(m, y_R)$ is the coefficient of the term with P_R^j factor. Otherwise, if l_i's have a unique minimum l, then since $Q_l(P_R) = 0$ for all $p \in (0,1)$, letting $p \to 1$ shows that $g(m, y_R) = 0$ where $g(m, y_R)$ is the coefficient of the term with $(1-P_R)^l$ factor. Whichever coefficient is 0, we remove that term from $h(P_R)$ before the next step. Starting with $j = 1$ and $l = 1$, we continue this procedure until all the terms in Equation 18 are removed, concluding that $g(m, y_R) = 0$ for all (m, y_R)

Since (m, y_{Rm}) is the complete and sufficient statistic and $\hat{P}_{R1} = X_{R1}/n_1$ is unbiased, by the Rao-Blackwell theorem the UMVUE of P_R is given by $\widetilde{P_R} = E\{X_{R1}|(m, y_{Rm})\}$. If $m = 1$, we have $\widetilde{P_R} = \hat{P}_{R1}$, but if $2 \leq m \leq K$, then the conditional probability mass function of X_{R1} given (m, y_R) in ω is

$$\frac{Pr\{X_{R1} = x_{R1}, M = m, Y_{Rm} = y_{Rm}\}}{Pr\{M = m, Y_{Rm} = y_{Rm}\}}$$

$$= \frac{Pr\{X_{R1} = x_{R1}, Y_{Rm} = y_{Rm}, a_k \leq y_{Rk}, k = 2, \ldots, m-1|\mathbf{P}\}}{f(m, y_{Rm}|\mathbf{P})}$$

$$= \frac{\sum_{R(m,y_{Rm}|x_{R1})} \binom{n_1}{x_{R1}} P_R^{x_{R1}} (1-P_R)^{n_1 - x_{R1}} \cdots \binom{n_m}{x_{Rm}} P_R^{x_{Rm}} (1-P_R)^{n_m - x_{Rm}}}{f(m, y_{Rm}|\mathbf{P})}$$

where the summations in the numerator are over

$$R(m, y_{Rm}|x_{R1}) = \{(x_{R2}, \cdots, x_{Rm}) : x_{R2} + \cdots + x_{Rm} = y_{Rm} - x_{R1},$$

$$y_{Rk} \geq R_k + 1, y_{Tk} \leq T_k - 1, k = 1 \cdots (m-1)\}.$$

The conditional probability simplifies to

$$\frac{\binom{n_1}{x_{R1}} \sum_{R(m,y_{Rm}|x_{R1})} \binom{n_2}{x_{R2}} \cdots \binom{n_m}{x_{Rm}}}{\sum_{R(m,y_{Rm})} \binom{n_1}{x_{R1}} \cdots \binom{n_m}{x_{Rm}}}.$$

Therefore,

$$\tilde{P_R} = \frac{E\{X_{R1}|(m, y_R)\}}{n_1} = \frac{\sum_{x_{R1}} X_{R1}\binom{n_1}{x_{R1}}\sum_{R(m,y_{Rm}|x_{R1})}\binom{n_2}{x_{R2}}\cdots\binom{n_m}{x_{Rm}}}{n_1\sum_{R(m,y_{Rm})}\binom{n_1}{x_{R1}}\cdots\binom{n_m}{x_{Rm}}}$$
$$= \frac{\sum_{R(m,y_{Rm})}\binom{n_1-1}{x_{R1-1}}\binom{n_2}{x_{R2}}\cdots\binom{n_m}{x_{Rm}}}{\sum_{R(m,y_{Rm})}\binom{n_1}{x_{R1}}\cdots\binom{n_m}{x_{Rm}}}$$

where the summations in the numerator and the denominator are over

$$R(m, y_{Rm}) = \{(x_{R1}, \cdots, x_{Rm}) : x_{R1} + \cdots + x_{Rm} = y_{Rm},$$
$$y_{Rk} \geq R_k + 1, y_{Tk} \leq T_k - 1, k = 1 \cdots (m-1)\}.$$

5.3 Simulations

A series of simulations was constructed to compare the bias and the relative efficiency of the proposed estimator to the maximum likelihood estimator (MLE). The relative efficiency is the ratio of the mean squared error of the MLE to the variance of the proposed estimator. We considered three different clinical trial designs that include both response and toxicity; the Bryant and Day (1995) design, the Ray and Rai (2011b) design, and the Simon 2-Stage design combined with continuous toxicity monitoring (CTM) (Ivanova et al., 2005; Ray and Rai, 2011a). The Bryant and Day methodology is a two-stage design that considers both response and toxicity at the same time. The Ray and Rai design is able to examine both response ant toxicity. The toxicity examination is performed four times while the response is only examined twice. The final design we consider is a Simon 2-Stage design combined with a continuous toxicity monitoring originally proposed by Ivanova et al. (2005) but expanded by Ray and Rai (2011a).

An effort was made to ensure the parameters input into the various clinical trial designs were similar while considering the required sample sizes. The design parameters used in the simulations are listed in Table 9. Each design requires slightly different specification of the input criteria. For instance, the Bryant and Day design requires specification of the unacceptable "non-toxicity" rate under the null and the acceptable "non-toxicity" rate under the alternative. Table 9 reports these as $(1 - P\{\text{non-toxicity}\})$, which is the probability of toxicity under the null and alternative hypothesis. In order to simplify the discussion of the simulation results, the designs will be grouped based on the response rates used to create the design:

- Design I - $(P_{R0}, P_{RA}) = (0.10, 0.30)$
- Design II - $(P_{R0}, P_{RA}) = (0.20, 0.40)$
- Design III - $(P_{R0}, P_{RA}) = (0.30, 0.50)$.

TABLE 9

Design Parameters

Methodology	Design	P_{R0}	P_{RA}	P_{T0}	P_{TA}	α_R	α_T	α	β
Bryant &	I	0.10	0.30	0.40	0.25	0.05	0.05	.	0.10
Day	II	0.20	0.40	0.40	0.25	0.05	0.05	.	0.10
	III	0.30	0.50	0.40	0.25	0.05	0.05	.	0.10
Ray &	I	0.10	0.30	0.33	0.25	.	0.05	0.05	0.10
Rai	II	0.20	0.40	0.33	0.25	.	0.05	0.05	0.10
	III	0.30	0.50	0.33	0.25	.	0.05	0.05	0.10
Simon 2	I	0.10	0.20	0.33	.	0.05	0.05	.	0.10
Stage	II	0.20	0.40	0.33	.	0.05	0.05	.	0.10
with CTM	III	0.30	0.50	0.33	.	0.05	0.05	.	0.10

The simulations were repeated for many different values of response, which varied from 0 to $P_{RA} + 0.10$ by 0.01 increments. The toxicity rate was always set in the alternative region, since stopping early for toxicity will produce similar results to stopping early for futility. The simulation procedure for each design was repeated three times to account for three different values of correlation between toxicity and response. We consider the smallest possible correlation, a correlation of 0, and the maximum possible correlation. The effect of the correlation on the simulation results is negligible, so only the results associated with independence are reported here. The critical values and sample sizes are reported in Table 10. Each simulation was repeated 10,000 times where one simulation is an executed clinical trial.

The proposed estimator is similar to the Jung and Kim (2004) estimator when toxicity and response are monitored at the same time. It also important to note that it is not possible to use the proposed estimator in the design with the continuous toxicity monitoring. The methodology monitors the toxicity after each patient is enrolled. or $n_i = 1$ for i in $1, 2, \ldots, n$. The calculations were performed as if the trial was only a two stage design. So if the trial stopped before the first response evaluation, then the stage is set to 1 in the expression for the point estimate. If the trial stopped before the end of the trial but after the first Simon 2-Stage boundary, then the stage was considered the second stage.

Figure 16 displays the bias of the MLE over the range of possible response rates, which includes P_{R0} and P_{RA}. The bias of the MLE is the largest near P_{R_0} instead of near the extreme values 0 or $P_{R_A} + 0.10$. We can also see the additional stopping for toxicity induced by the Ray and Rai design creates a slightly different pattern in the bias of the MLE.

TABLE 10

Sample Sizes and Critical Values

Methodology	Design	Sample Sizes		Response Boundaries		Toxicity Boundaries	
		n_1	n	r_1	r	t_1	t_2
Bryant & Day	I	41	100	5	14	26	67
	II	37	103	8	27	23	69
	III	39	95	12	35	24	64

		$n_1 = n_2 = n_3 = n_4$	n	r_1	r	t_1	t_2	t_3	t_4
Ray & Rai	I	11	44	3	7	7	13	20	22
	II	13	52	6	13	9	19	19	24
	III	14	56	9	22	10	16	25	25

		n_1	n	r_1	r	$t_1, t_2, \ldots, t_n$
Simon 2 Stage with CTM	I	18	35	2	6	See Table 11
	II	19	54	4	15	See Table 11
	III	24	63	8	24	See Table 11

The absolute value of the bias of the MLE is the largest when it is used with the Simon 2-Stage design combined with continuous toxicity monitoring.

Figure 17 depicts the bias of the proposed estimator, which includes the modified estimator utilized for the Simon 2-Stage design with continuous toxicity monitoring. In each case, we can see there is not an obvious pattern. The bias is near 0 in each simulation. The maximum reduction in bias is approximately 90%. Ignoring the toxicity does not seem to affect the bias of the estimator very much.

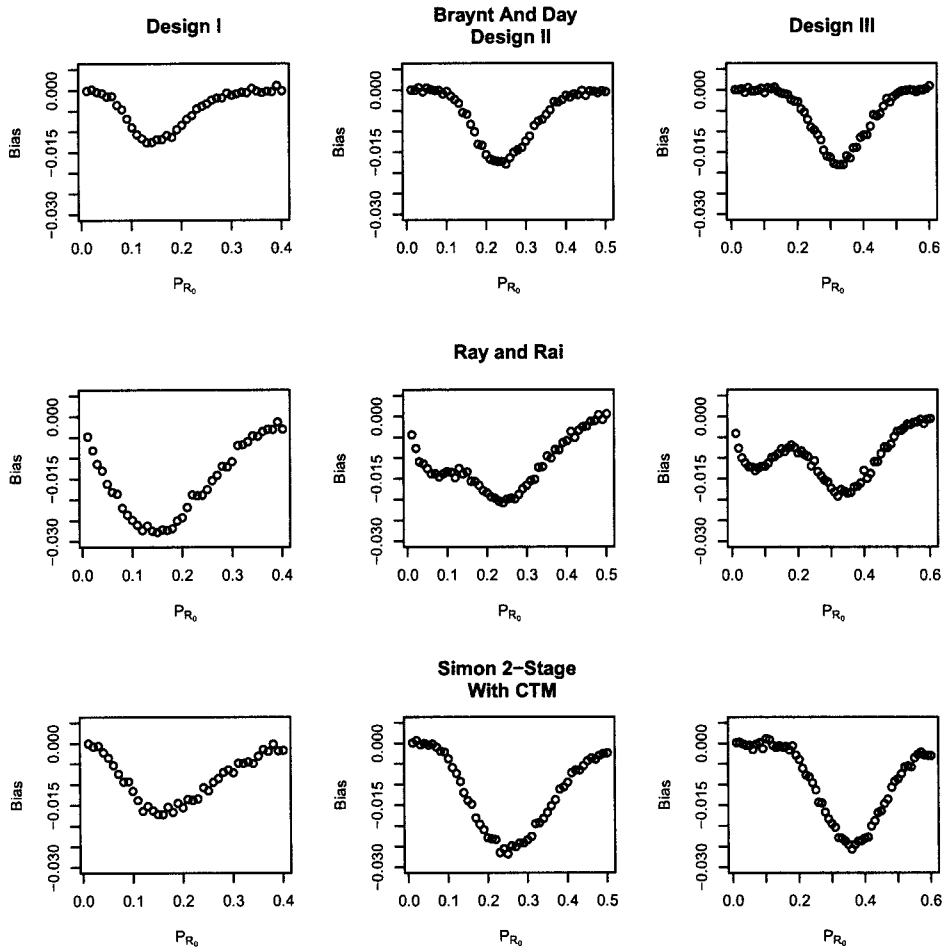

Figure 16. Bias of Maximum Likelihood Estimate

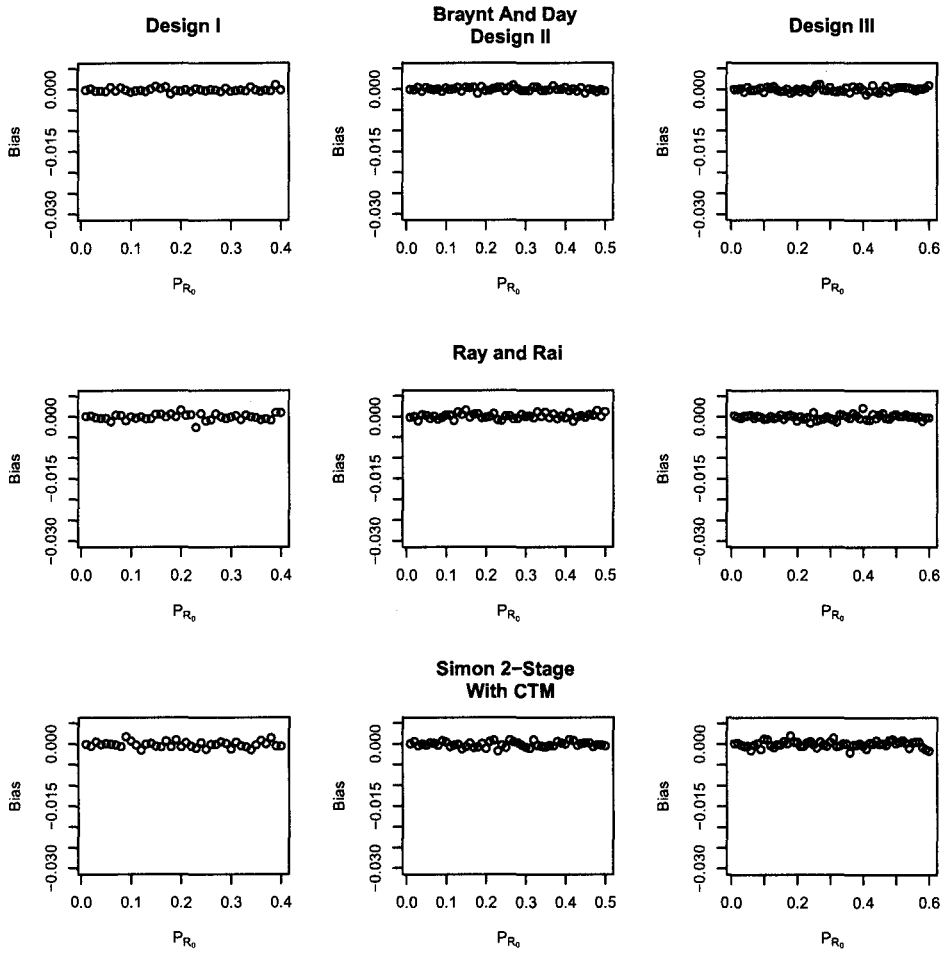

Figure 17. Bias of Proposed (or Modified) Estimator

The relative efficiency of the MLE to the proposed estimator is pictured in Figure 18. The MLE is more efficient than the new estimator near the value P_{R0} in the Bryant and Day, as well as the Simon 2-Stage design combined with the continuous toxicity monitoring. The proposed estimator is more efficient in all situations when the response rate is larger than P_{R0}, but not quite as large as P_{RA}. The MLE is the most efficient when the responses rates are small in the Ray and Rai design. The curve is also bimodal, in two of the three general design categories, utilizing this methodology.

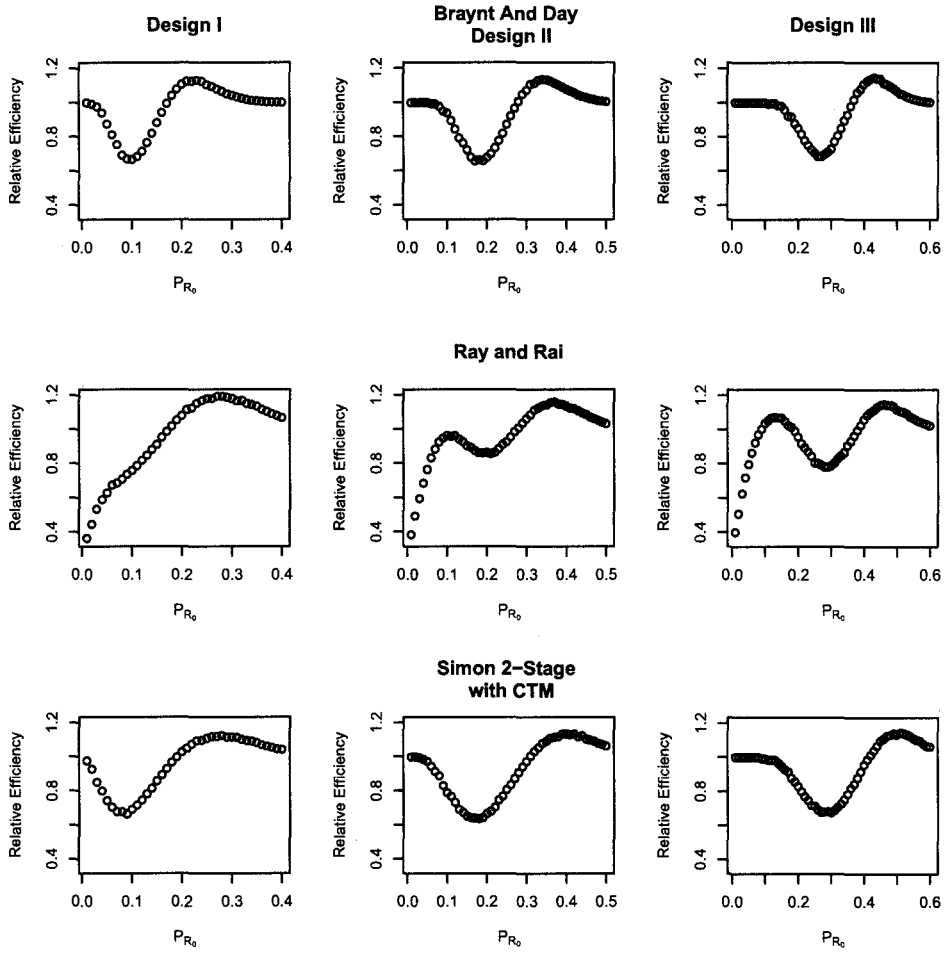

Figure 18. Relative Efficiency of the Maximum Likelihood Estimate to Proposed (or Modified) Estimator

The performance of the modified estimator in the continuous toxicity monitoring situation is good. The bias is small across the range of response rates. Also, the proposed estimator is more efficient than the MLE over portions of the range of the response rates. Next, we consider the benefit of incorporating the additional information from the toxicity monitoring into the estimator. Figure 19 graphs the relative efficiency of the proposed estimator to the modified version of the estimator in the Ray and Rai design. The relative efficiency is examined through the ratio of the mean squared error of the modified estimator to the variance of the proposed estimator. We can see for response rates larger than 0.10 the proposed estimator is more efficient than the modified version that only incorporates information from two of the four stages. Thus, the new estimator is more efficient but one could also use the modified version since it reduces the bias when compared

to the MLE.

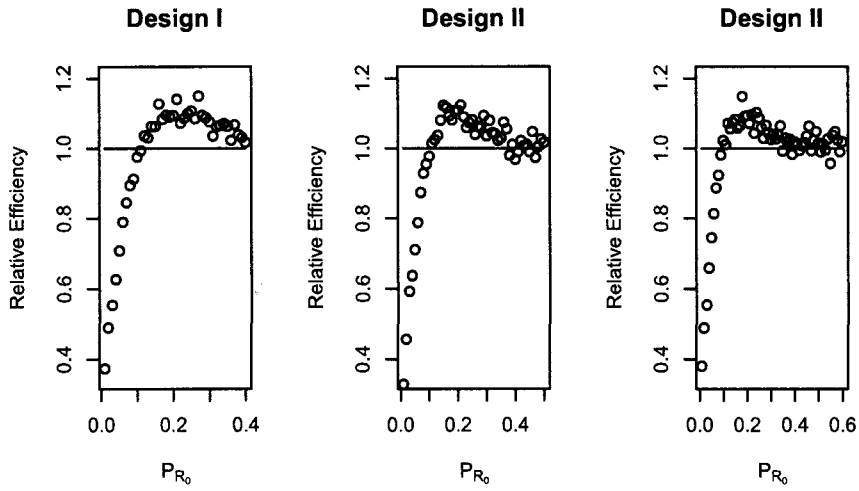

Figure 19. Relative Efficiency of Modified Estimator to Proposed Estimator in the Ray and Rai Design

5.4 Discussion

The investigation was conducted to determine the best estimator to use after a clinical trial that leveraged the Simon 2-Stage design with the continuous toxicity monitoring methodology. We found an estimator that can be utilized in bivariate designs that stop early for either response or toxicity. We also showed that the new estimator is the uniformly minimum variance unbiased estimator. We then modified the estimator so it could be used with continuous toxicity monitoring. Through simulation we discovered that it is also an unbiased estimator, but the proposed estimator is more efficient over specific ranges of the response rate.

TABLE 11

Continuous Toxicity Monitoring Boundaries - Design I, II, and III

Design I $P_{T0} = 0.33$, $\alpha_T = 0.05$, $n = 35$			Design II $P_{T0} = 0.33$, $\alpha_T = 0.05$, $n = 54$			Design III $P_{T0} = 0.33$, $\alpha_T = 0.05$, $n = 63$		
Minimum # Subjects	Maximum # Subjects	# of Subjects with a Toxicity (t_i)	Minimum # Subjects	Maximum # Subjects	# of Subjects with a Toxicity (t_i)	Minimum # Subjects	Maximum # Subjects	# of Subjects with a Toxicity (t_i)
4	4	4	4	4	4	4	6	5
5	5	5	5	5	5	7	7	6
6	7	6	6	7	6	8	8	7
8	9	7	8	9	7	9	11	8
10	12	8	10	11	8	12	12	9
13	14	9	12	13	9	13	15	10
15	16	10	14	15	10	16	17	11
17	18	11	16	17	11	18	20	12
19	20	12	18	20	12	21	22	13
21	22	13	21	22	13	23	24	14
23	24	14	23	24	14	25	26	15
25	27	15	25	26	15	27	28	16
28	29	16	27	29	16	29	31	17
30	31	17	30	31	17	32	33	18
32	32	18	32	33	18	34	35	19
33	35	19	34	35	19	36	38	20
			36	38	20	39	40	21
			39	40	21	41	42	22
			41	43	22	43	45	23
			44	44	23	46	47	24
			45	47	24	48	50	25
			48	50	25	51	52	26
			51	52	26	53	54	27
			53	54	27	55	57	28
						58	59	29
						60	61	30
						62	63	31

CHAPTER 6

PHASE IIB OR III CLINICAL TRIAL DESIGNS THAT INCLUDE MULTIPLE ENDPOINTS

The phase III or IIb clinical trial is typically a large randomized study intended to confirm the efficacy observed earlier in the treatment development process. The patients are randomized into at least two arms, but the designs may include three or more arms as well. The designs often include provisions for early examination of the data using group sequential theory. The early methodologies by Pocock (1977), as well as O'Brien and Fleming (1979), allow early termination of the trial if there was evidence to reject the null hypothesis. Other authors, such as Emerson and Fleming (1989), consider group sequential procedures that allow the trial to terminate early in favor of or to reject the null hypothesis. Jennison and Turnbull (1999) refer to these as "inner-wedge" designs since the futility stopping boundaries form a wedge inside of the efficacy boundaries. The designs focus on univariate normally distributed variables, such as clinical response or event free survival.

Halperin et al. (1982) propose a method that allows a study to terminate early for futility based on stochastic curtailment, which is a flexible approach to monitor the emerging results of a clinical trial. Lachin (2005) describes stochastic curtailment as a decision to terminate the trial based on an assessment of the conditional power (CP). CP is the conditional probability that the final result will exceed the critical value given the accrued data and an assumption about the data to be observed during the remainder of the study. Ying and Clarke (2010) outline a flexible time-varying conditional power boundary methodology, which allocates portions of the type II error over time based on the typical alpha-spending functions. The methodology has similar benefits including the ability to pre-specify the interim analysis. One could also use the flexibility associated with the alpha-spending approach to modify the exact timing of the interim analysis. The resulting methodology is an intuitive expansion and application of the conditional power approach to futility monitoring.

The safety of the clinical trial subjects is an important endpoint that must be considered in all clinical trials. According to the ICH E9 guidelines on Statistical Considerations in Clinical

Trials, safety must be monitored in all clinical studies (Chow and Liu, 2004). The studies should incorporate procedures that allow for early termination of the trial for safety reasons. There are two different types of bivariate test procedures which can include multiple stages. One could formulate a test procedure which uses a global test statistic such as Hotelling's t-test (Hotelling, 1931). It is also possible to construct a test procedure that considers the marginal endpoints separately. Jennison and Turnbull (1993), as well as Cook and Farewell (1994), propose bivariate test procedures that consider both endpoints separately, but attempts to control the global type I and II error rates. The group sequential procedures allow one to evaluate both efficacy and futility, as well as patient safety, through the conduct of the trial. There are also more sophisticated methods proposed to control the average type I error rate as well. Chuang-Stein et al. (2007) suggest a method that attempts to control the average type I error rate resulting in slightly different significant levels associated with the marginal hypothesis test. In this chapter, we will consider combining the conditional power approach with the customary method proposed by Jennison and Turnbull (1993). Currently, it is difficult to allow early termination for only futility or safety considerations.

The fixed sample size design described by Jennison and Turnbull (1993) examines both endpoints separately but simultaneously. The usual critical value, Z_α, is used to evaluate both endpoints in separate tests where both null hypothesis much be rejected to declare the treatment successful. The sample size required to achieve the power, $1-\beta$, relies on the joint bivariate normal (BVN) distribution, including the correlation between the test statistics. The overall type I error rate maybe much smaller than expected, resulting in a much more conservative trial than desired, especially in the phase IIb clinical trial setting. In Section 6.1, we describe the fixed sample size test in detail including the potential issue with the overall type I error rate. We then propose a modification which searches for a critical value, C, and corresponding sample size that controls the overall type I and II error rates, respectively. In Section 6.2, we combine both fixed sample size methodologies with the time-varying conditional power concept which allows us to stop the trial early for either futility or safety. The two fixed sample size methods are examined through simulation in Section 6.4. The simulation results associated with the conditional power approach combined with the fixed sample size methodologies is discussed in Section 6.5. The conditional power combined with the customary fixed sample size approach, that utilizes the critical value, Z_α, during the early examinations of the data, inflates the type II error rate. It is found that the type II error rate of the customary procedure is controlled if the critical value, C, is leveraged for the interim analysis and the critical value, Z_α, is utilized in the final analysis. A final discussion of the results is included in Section 6.6.

6.1 Fixed Sample Design

First, we will consider the fixed sample size situation in which subjects can be randomized into either a treatment group or a control group. The hypothesis to be evaluated is

$$\begin{aligned} H_0 : P_{RT} \leq P_{RC} \quad &\text{or} \quad P_{TT} \geq P_{TC} + \delta_T \\ &\text{versus} \\ H_A : P_{RT} > P_{RC} \quad &\text{and} \quad P_{TT} < P_{TC} + \delta_T \end{aligned} \tag{19}$$

where P_{RT} is the response rate from the treatment group, P_{RC} is the response rate from the control group, P_{TT} is the toxicity rate from the treatment group, P_{TC} is the toxicity rate from the control group, and δ_T is the increase in toxicity allowed for the new treatment to achieve the response rate. The hypothesis test displayed in Equation 19 can also be broken down into the marginal hypotheses

$$\begin{aligned} H_{10} : P_{RT} \leq P_{RC} \quad &\text{versus} \quad H_{1A} : P_{RT} > P_{RC} \\ &\text{and} \\ H_{20} : P_{TT} \geq P_{TC} + \delta_T \quad &\text{versus} \quad H_{2A} : P_{TT} < P_{TC} + \delta_T \end{aligned} \tag{20}$$

where both H_{10} and H_{20} must be rejected to declare the new treatment successful. Jennison and Turnbull (1993), as well as Cook and Farewell (1994), propose methods that control the type I error rate for the marginal tests. Then, the procedure searches for a sample size that meets the power requirements.

Let $\overline{\mathbf{X_T}} = (\overline{x_{RT}}, \overline{x_{TT}})$ be the mean response and toxicity vector for the treatment group and $\overline{\mathbf{X_C}} = (\overline{x_{RC}}, \overline{x_{TC}})$ be the mean response and toxicity vector for the control group. Let s_{RT} and s_{RC} be the sample standard deviations for the response measurement from the treatment and the control groups, respectively. Also, let s_{TT} and s_{TC} be the sample standard deviations for the toxicity measurement from the treatment and control groups. Let $\mathbf{Z_{RT}} = (Z_R, Z_T)$ be the standardized bivariate test statistic where

$$Z_R = \frac{\overline{x_{RT}} - \overline{x_{RC}}}{\sqrt{\frac{s_{RT}^2}{n} + \frac{s_{RC}^2}{n}}}$$

and

$$Z_T = \frac{\overline{x_{TC}} - \overline{x_{TT}} + \delta_T}{\sqrt{\frac{s_{TT}^2}{n} + \frac{s_{TC}^2}{n}}}$$

Note that $\mathbf{Z_{RT}}$ follows a $BVN(0, \Sigma)$, under the null hypothesis, where

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Response and toxicity can both be measured as correlated binomial random variables, which have asymptotic normal distributions.

Jennison and Turnbull (1993) control the marginal type I error rates. If the marginal type I error rates are controlled at the same α-level, say 0.05, then the true type I error rate of the combined test maybe be much smaller than expected. Under the null hypothesis assuming the standardized tests statistics are independent,

$$\begin{aligned}
P\{Z_R > Z_\alpha \ \cap \ Z_T > Z_\alpha | H_0\} & \\
&= P\{Z_R > Z_\alpha | H_0\} P\{Z_T > Z_\alpha | H_0\} \\
&= \frac{P\{Z_R > Z_\alpha \cap H_0\}}{P\{Z_R > Z_\alpha\}} \frac{P\{Z_T > Z_\alpha \cap H_0\}}{P\{Z_T > Z_\alpha\}} \\
&= \frac{P\{Z_R > Z_\alpha \cap (H_{10} \text{ or } H_{20})\}}{P\{Z_R > Z_\alpha\}} \frac{P\{Z_T > Z_\alpha \cap (H_{10} \text{ or } H_{20})\}}{P\{Z_T > Z_\alpha\}} \\
&= \frac{P\{Z_R > Z_\alpha \cap H_{10}\}}{P\{Z_R > Z_\alpha\}} \frac{P\{Z_T > Z_\alpha \cap H_{20}\}}{P\{Z_T > Z_\alpha\}} \\
&= P\{Z_R > Z_\alpha | H_{10}\} \cdot P\{Z_T > Z_\alpha | H_{20}\} \\
&= 0.05 \cdot 0.05 = 0.0025.
\end{aligned}$$

The type I error rate also depends on the correlation and can vary from 0 to 0.05 as the correlation varies from -1 to 1. It is also possible to consider the more general case

$$P\{Z_R > Z_{\alpha R} \ \cap \ Z_T > Z_{\alpha T} | H_0\}$$

where $Z_{\alpha R}$ is associated with response and $Z_{\alpha T}$ is associated with toxicity.

Although the marginal hypotheses maybe of interest in some settings, it is also possible that the overall type I error rate is too conservative in other instances. Using the numeric integration algorithm developed by Genz (2004), it is possible to search for two critical values, C_R and C_T, such that

$$P\{Z_R > C_R \ \cap \ Z_T > C_T | H_0, \rho\} \leq \alpha.$$

In order to reduce the complexity of the discussion, we will consider the case when $C_R = C_T = C$ and only search for a single critical value, C, that satisfies

$$P\{Z_R > C \ \cap \ Z_T > C | H_0, \rho\} \leq \alpha.$$

The approach will increase the type I error rate associated with the marginal hypothesis test, but will control the overall type I error rate at the specified level for the joint hypothesis. A bisection algorithm is used to search for C and the concept could be expanded to more than two endpoints. It is also important to note that Zhao, Grambsch, and Neaton (2005) evaluated several bivariate

integration algorithms and the Genz algorithm was preferred, since it is accurate and executes quickly.

Once the critical value, C, is determined, then we can search for the sample size that meets the required type II error rate. The probabilities under consideration are

$$P\{Z_R > C \cap Z_T > C | H_A, \rho\} \geq 1 - \beta.$$

The specific double integral under consideration is

$$\int_{-\infty}^{\left(C - \frac{\sqrt{n}\delta_R}{\sqrt{P_{RT}\cdot(1-P_{RT})+P_{RC}\cdot(1-P_{RC})}}\right)} \int_{-\infty}^{\left(C - \frac{\sqrt{n}\delta_T}{\sqrt{P_{TT}\cdot(1-P_{TT})+P_{TC}\cdot(1-P_{TC})}}\right)} f(Z_R, Z_T) dZ_T \, dZ_R$$

where

$$f(Z_R, Z_T) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(\frac{-(Z_R^2 - 2\rho Z_T Z_R + Z_T^2)}{2(1-\rho^2)}\right).$$

The fixed sample size design will be evaluated in the Section 6.3.

6.2 Futility and Safety Monitoring

The ability to stop a trial early for futility is extremely important since it prevents future patients from receiving an ineffective treatment. The conditional power approach is a popular methodology that allows one to predict the ability to reject the null hypothesis given the current data along with an assumption about the detect size (Lachin, 2005). Ying and Clarke (2010) note that the B-value is often used as a data monitoring tool, which is a function of the sequence of standardized test statistics $\{Z_1, Z_2, \ldots, Z_K\}$ calculated at information times $t_i = n_i/N$ for $i = 1, 2, \ldots, K$. The B-value at information time t_i is typically defined to be $B(t_i) = Z_i t_i^{\frac{1}{2}}$. The conditional power at information time t_i, assuming no efficacy analysis, is approximated by

$$\text{CP}(Z_i) = 1 - \Phi\left(\frac{b(1) - B_i(t_i) - (1-t_i)\Theta}{\sqrt{(1-t_i)}}\right) \tag{21}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and Θ is the drift parameter. The conditional power is typically calculated with the desired detect size and compared to a fix γ, such as 0.5. The trial is stopped early for futility if $\text{CP}(Z_i) < \gamma$.

Ying and Clarke (2010) propose a flexible method to create a time-varying conditional power boundary denoted as $\gamma_i(t_i)$, for $i \in \{1, 2, \ldots, K\}$. The formulation relies on the following observations associated with a K interim analysis performed at information times $\{t_1, t_2, \ldots t_K\}$. Before the final analysis, $\gamma_i(t_i)$ is determined such that

$$P\{\text{CP}_1(\Theta) \geq \gamma_1, \ldots, \text{CP}_{i-1}(\Theta) \geq \gamma_{i-1}, \text{CP}_i(\Theta) < \gamma_i) = f(t_i) - f(t_{i-1}) \tag{22}$$

for $i = 1, 2, \ldots K$ where f is an increasing function with $f(0) = 0$ and $f(1) = \beta$ and

$$\gamma_i(t_i) = \Phi\left(C_i\sqrt{\frac{t_i}{1-t_i} + \frac{Z_{1-\beta}}{\sqrt{1-t_i}}}\right). \tag{23}$$

The conditional power formulation in Equation 22 can be equivalently expressed in terms of the standardized test statistics

$$P\{Z_1 - \Theta\sqrt{(t_1)} \geq C_1, \ldots, Z_{i-1} - \Theta\sqrt{(t_{i-1})} \geq C_{i-1}, Z_i - \Theta\sqrt{(t_i)} < C_i\} = f(t_i) - f(t_{i-1})$$

for $i = 1, 2, \ldots K$. The authors note that $\{Z_t\sqrt{t} - \Theta t : 0 \leq t \leq 1\}$ follows a Brownian motion process and thus apply the usual alpha-spending functions to the type II error rate to determine the futility monitoring constants $\{C_1, C_2, \ldots, C_K\}$.

The same approach can be applied to the marginal hypothesis tests H_{10} and H_{20} described in Equation 21, since both must be rejected in order to reject the null hypothesis. If either one of the two hypotheses appears to not be able to reject the null hypothesis, then the study should be stopped early. In the formulation of the hypothesis H_{20}, the same concept can be applied to the monitoring of toxic events. If it does not appear that the number of toxic events will be less than the number experienced on the control treatment plus the margin, then the trial can be stopped early as well. Section 6.3 will explore the type I and II error rates resulting from the combination of the conditional power approach and the bivariate test procedure. We will also consider two different alpha-spending functions.

6.3 Simulations

First, we will consider the fixed sample size design discussed in Section 6.1. We will consider two different methods to determine the fixed sample size. Method 1 is similar to the method described by Jennison and Turnbull (1993) and will use the typical critical values, Z_α, for the marginal tests. Then we will use the bivariate normal distribution to search for the sample size based on the required power. Method 2 will search for the critical values, C, such that the overall type I error rate is controlled at the specified level. Then the procedure will search for the sample size that controls the type II error rate. We will examine the effects of the correlation on the type I and II error rates when it is different than assumed during the design of the trial. The second set of simulations will combine Method 1 with the conditional power concept on the marginal hypothesis test discussed in Equation 21. We will also consider the conditional power methodology for efficacy and safety combined with Method 2. Then we will examine the type I and II error rates, as well as the probability of stopping early under different circumstances. The values

reported in the subsequent tables are based on 100,000 simulations, or each error rate reported is an average based on 100,000 simulated clinical trials.

6.4 Simulation - Fixed Sample Size

The clinical trial utilized in the simulation is designed to detect a 0.20 increase in response rate, as well as a toxicity rate that is no more than 0.20 larger than the control. For proposes of the simulation and calculations, we assume a response rate of 0.30 and toxicity rate 0.30 for the standard treatment. We will set the marginal type I error rates at 0.05 and the overall power at 0.80. The only way we can make a type I error under the null is if we reject the null hypothesis for both endpoints. The overall type I error rate for Method 2 will be fixed at 0.05, and then we will search for the critical value that produces the result. The current setup assumes equal allocation of subjects between the treatment and the control.

Table 12 contains the sample sizes and critical values required to control the type I and II error rates for the specific hypothesis described above. We can see that the sample size decreases as ρ increases in Method 1. We can also see that the critical values, C, and sample sizes increase in Method 2 as the correlation increases.

TABLE 12

Sample Sizes and Critical Values

	Method 1		Method 2	
Correlation	Critical Value	Sample Size	Critical Value	Sample Size
-80%	1.645	190	0.163	46
-60%	1.645	190	0.334	58
-40%	1.645	188	0.485	70
-20%	1.645	188	0.625	80
0%	1.645	186	0.760	90
20%	1.645	182	0.894	100
40%	1.645	178	1.030	110

Once the critical values and the sample sizes are calculated with a specified ρ, then we consider the effect of incorrectly specifying the correlation on the type I and II error rates. Table 13 contains the simulated power based on different values of ρ when independence is assumed to design the trial. The correlation will not take on the full range of [-1, 1] in this instance due to the underlying multinomial distribution associated with the various combinations of response and toxicity.

We can see that the power increases as ρ increases from the minimum value of -80%. We can also see that power is achieved when $\rho = 0$ but is less than desired when the correlation is negative.

TABLE 13

Simulated Power and Type I Error Rate Under Various Correlations When Assumed to be Independent

	Method 1		Method 2	
Correlation	Type I Error Rate	Power	Type I Error Rate	Power
-80%	0.000%	79.611%	0.004%	78.639%
-60%	0.001%	79.724%	0.542%	78.966%
-40%	0.014%	79.814%	1.603%	79.102%
-20%	0.091%	80.088%	2.955%	79.350%
0%	0.221%	80.670%	4.556%	80.011%
20%	0.478%	81.544%	6.431%	80.895%
40%	0.915%	82.309%	8.622%	81.812%

The effect of the correlation on the type I error rate is also available in Table 13. In general, we can see that the simulated type I error rate for Method 1 is much smaller than for Method 2. We can also see that the type I error rate increases as ρ increases in both methods. The simulated type I error rate exceeds the specified alpha-level in Method 2, but achieves a maximum value of 0.915% in Method 1.

The final item to consider is the simulated type I error rate when we reject one of the two hypotheses, but fail to reject the other. For instance, does the type I error rate increase if we reject H_{10} but fail to reject H_{20}. Table 14 contains the results of the simulation when we reject one of the hypotheses but fail to reject the other. We can see that the resulting type I error rates are all less than 5% for Method I but nearly 20% for Method 2.

TABLE 14

Type I Error Rate When One Endpoint Falls in the Rejection Region and The Other Does Not

	Hypothesis Rejected	Hypothesis Not Rejected	Type I Error Rate	Standard Error
Method 1	H_{10}	H_{20}	4.73%	0.067%
Method 2	H_{10}	H_{20}	18.68%	0.123%
Method 1	H_{20}	H_{10}	3.94%	0.062%
Method 2	H_{20}	H_{10}	19.39%	0.125%

Both designs can be used to examine response and toxicity at the same time. The reduced sample size in Method 2 creates an additional cost associated with the type I error rate, if we are able to reject only one of the two hypothesis. Next we will add in monitoring that allows the trial to terminate early for either efficacy or safety.

6.5 Simulation - Conditional Power

The time varying conditional power approach requires one to specify an alpha-spending function, the type II error to be spent, and the number of interim analysis. In the simulations we investigate the Pocock and the O'Brien-Fleming type boundaries combined with Method 1 and Method 2. We consider 3 examinations before the end of the trial that allow one to stop early for futility or safety. At the end of the trial we only allow rejection of the null hypothesis. Rejection requires both H_{10} and H_{20} to be rejected in order to declare the treatment successful.

The fixed sample sizes associated with independence are used in the simulations. The simulations are designed to repeat each clinical trial 100,000 times. Initial investigations demonstrated an inflation of the type II error rate, if we created the conditional boundary values based on the full power. The conditional boundary values are constructed with 50% of the specified type II error, and then it is equally allocated to each end point. This is similar to a Bonferroni correction in the usual multiplicity problem. The result is boundary values based on a type II error rate of 5% per endpoint point.

The decision rules applied to the i^{th} analysis for $i = 1, 2, 3, 4$ are:

1. At time t_i for $i < 4$, calculate CP(Θ) based on the accumulated data and the desired effect size for each endpoint. If CP(Θ) $\leq \gamma_i$ for either end point, then stop the trial for either safety or futility;

2. else, continue the trial and repeat step 1 for $i < 4$;

3. at the final analysis, if $Z_R > C$ and $Z_T > C$, where C is the appropriate critical value, then we declare the treatment successful; otherwise we cannot reject the null hypothesis.

Table 15 provides the conditional power boundary values used in the simulation. We can see that the Pocock type boundary is much larger early in the trial than the O'Brien-Fleming type design. This is similar to the alpha-spending approach, since the Pocock type boundary will have a much higher probability of stopping very early when compared to the O'Brien-Fleming boundary values.

The probability of stopping at each stage is displayed in Table 16. We can see that the probability of stopping early under the null hypothesis is very high in Method 1 for both the

TABLE 15

Conditional Power Boundary Values

Boundary Type	t_1	t_2	t_3
O'Brien-Fleming	0.1906	0.1709	0.1052
Pocock	0.4292	0.2113	0.0386

0'Brien-Fleming and the Pocock type boundaries. The probability of stopping early under the alternative is also very large for both boundary value types, which will increase the overall type II error rates. The simulated type II error rates are reported in Table 17. We can see the simulated type II error rate is 29.16% with the O'Brien-Fleming boundary values and 60.43% with the Pocock boundary value types. It appears aggressive monitoring using the conditional power approach increases the type II error rate.

TABLE 16

Percentage of Trials That Stopped Early

		Under the Null Hypothesis			Under the Alternative Hypothesis		
	Stage	Method 1	Method 1a	Method 2	Method 1	Method 1a	Method 2
O'Brien-Fleming	1	67.48%	5.63%	6.60%	9.55%	0.08%	0.46%
	2	25.67%	35.35%	34.84%	8.52%	0.50%	3.07%
	3	4.70%	31.00%	30.12%	2.57%	0.50%	3.69%
	Total	97.85%	71.98%	71.56%	20.64%	1.08%	7.22%
Pocock	1	96.98%	36.46%	36.49%	55.38%	2.79%	6.99%
	2	1.56%	22.02%	22.01%	1.41%	0.60%	2.71%
	3	0.48%	9.96%	11.98%	0.27%	0.12%	1.15%
	Total	99.02%	68.44%	70.48%	57.19%	3.51%	10.85%

TABLE 17

Simulated Type I and II Error Rates

		Method 1	Method 1a	Method 2
O'Brien-	Type I	0.182%	0.237%	4.57%
Fleming	Type II	29.16%	19.54%	19.37%
Pocock	Type I	0.150%	0.253%	3.91%
	Type II	60.43%	20.96%	23.04%

Method 2 utilizes the sample sizes and critical values from Table 12 corresponding with independence. In Table 16, we can see that the total probability of stopping early for either futility or safety is at most 71.56% under the null hypothesis. The Pocock type boundary values stop at the first analysis 36.46% while the O'Brien-Fleming type only stops at the first stage 6.60%. The percentage of clinical trials terminating early increases for the O'Brien-Fleming method but decreases for the Pocock method. The simulated type II error associated with the O'Brien-Fleming and Pocock boundary values are 19.37% and 23.04%, respectively. The Pocock type boundary also results in a slightly more conservative trial with a simulated type I error rate of 3.91% while the simulated type I error rate of the O'Brien-Fleming is 4.57%.

Method 1a is a modification of the calculation of the conditional power associated with Method 1. Method 1a replaces $b(1) = Z_\alpha$ in equation 20 with the critical value, C, from Table 12 associated with Method 2. The critical value, C, from Table 12 is the critical value used to evaluate the joint hypothesis. The probability of stopping early under both the null and alternative hypothesis for Method 1a is similar to Method 2. The simulated type I and II error rates reported in Table 17 are also similar to the error rates for Method 2.

It is possible to construct conditional boundary values in various ways but it is suggested that simulations be performed to evaluate the designs. For instance, we considered creating O'Brien-Fleming boundary values for Method 1 with marginal beta set at 2.5% which results in conditional power boundaries $\gamma_i = (0.1209, 0.0932, 0.0385)$. The resulting probability of stopping at Stage 1 is 39.65%, Stage 2 is 42.11%, and Stage 3 is 11.99%. The total percentage of stopping is 93.75%. The simulated type I error rate is 0.222% while the type II error rate is 22.67%. The results are similar to the type I and II error rates of the fixed sample size design. Only 10.04% of the clinical trials stopped early under the alternative hypothesis.

6.6 Discussion

The bivariate test procedure allows one to design a trial that considers both endpoints simultaneously while controlling the overall type I and II error rates. Method 1 controls the type I

error when one endpoint falls in the rejection region while the other endpoint does not, but also requires a larger sample size. Method 2 requires a smaller sample size, but the type I error rate will be inflated if one endpoint is rejected while the other is not. Also, the overall type I and II error rates will be affected by misspecification of the correlation. It is important to evaluate the sample sizes and critical values, under different assumptions, to understand the characteristics before they are used in practice.

The conditional power applied to both endpoints allows one to stop the trial early for either futility or safety. The choice of alpha-spending function and the amount of type II error rate to allocate drastically affects the overall type II error rate. The choice of alpha-spending function also affects the probability of stopping early under the null or alternative hypothesis. The choice also affects the overall type I and II error rates. Method 1C preserves the type I error rate if one of the hypotheses is rejected. It also preserves the type II error rate while allowing early examinations for both futility and safety. The method could be used with the ideas proposed by Chuang-Stein et al. (2007). It is important to note that the design should be evaluated through simulations before it is implemented in order to fully understand the operating characteristics.

In the end, the bivariate design fills a unique need that consider both endpoints simultaneously that controls the over type I and II error rates. The inclusion of the conditional power approach allows one to also examine the data early and make appropriate decisions protecting the patients.

CHAPTER 7

CONCLUSION

The requirement to monitor the safety of the clinical trial's subjects in the phase II and III studies motivated the dissertation research. Phase II and III clinical trials primarily focus on the efficacy of the new treatment. The trials are specifically designed to evaluate the response of the new drug treatment or the various survival rates. The safety of the participants cannot be neglected.

7.1 Concluding Summary

In practice, the clinical studies are designed to evaluate the primary endpoints. The toxicity considerations are included in an ad hoc manner outside of the formalized design. The motivating example designed a trial utilizing the same approach. First, the sample sizes and critical values were determined based on the Simon 2-Stage design. Then the toxicity was monitored after each patient was treated, utilizing the continuous toxicity monitoring methodology.

At the time the trial was designed, the operating characteristics of the combined procedure were unknown but required to evaluate the resulting design. Recursive expressions that accurately describe operating characteristics were discovered. The expressions assumed an underlying multinomial distribution. The operating characteristics include the type I and II error rates, the probability of early termination (PET), and the average sample size (ASN).

The theoretical expressions of the operating characteristics associated with the combined procedure unlocked many possibilities. First, we discovered the effect of the correlation between response and toxicity on the type I error rate. As the correlation increases the overall type I error rate decreases, or the combined procedure becomes more conservative. We also discovered an optimal choice for the type I error rate associated with the toxicity monitoring, or it should be set at 5%.

The decrease in the overall type I error rate of the combined procedure was disconcerting but a consequence of the ad hoc methodology. The combined procedure assumes that response and toxicity are independent, which may not necessarily be true. There is no way to fix the assumption

since the trial is designed in two steps. Therefore, the expressions for the operating characteristics were leveraged to create a flexible, bivariate clinical trial design that can monitor toxicity on a different schedule than response. The flexible design can include the simple bivariate two-stage design or the continuous toxicity monitoring methodology that examines response twice.

An example design that monitors toxicity four times and response twice, at the second and forth toxicity monitoring, was constructed and evaluated through simulation. The design is able to control the overall type I and II error rates. The type I error rate is inflated when one endpoint is in the null and the other is in the alternative. The issue is caused by the limited sample sizes expected in the phase II clinical trial setting. The result is also similar to other bivariate phase II clinical trial designs proposed by Conaway and Petroni (1995), as well as Bryant and Day (1995). Also, the type I error rate may be larger than expected if the true correlation is smaller than specified.

The next topic to consider was proper inference after the conduct of the bivariate, multiple-stage clinical trial. The maximum likelihood estimate (MLE) was found to be biased due to the multiple examinations of the response and toxicity. A uniformly minimum variance unbiased estimator (UMVUE) was discovered, which successfully removed the bias over the range of possible response rates. The point estimator had to be modified to work in the combined procedure, but the estimator preserved the reduction in bias. We also learned that the UMVUE is more efficient than the modified estimator, so the information associated with toxicity should not be ignored.

Finally, the natural extension of the research led to multiple-stage, bivariate clinical trials utilized in the phase IIb or III setting. The phase IIb/III clinical trials are typically multiple-arm and require larger sample sizes. The multivariate normal distribution is applicable in the comparative trial. Therefore, the methods proposed by Jennison and Turnbull (1999) were combined with the stochastic curtailment. The reverse multiplicity issue was addressed with a Bonferroni type correction to preserve the overall type II error rate. The stochastic curtailment also required a new critical value to be utilized. The new critical value prevented inflation of the type II error rate caused by the repeated hypothesis testing.

In summary, the toxicity considerations were formally incorporated into the clinical trial designs. Appropriate inference procedures were developed so that point estimation and other analytics can be performed after the conclusions of the studies. The research conducted to this point is an attempt to make clinical trial designs that formally include toxicity and response more appealing to biostatisticians, as well as principal investigators.

7.2 Future Research

The results presented previously expand the bivariate multiple-stage designs in an attempt to make them more appealing in practice. Although we have made many significant advancements, additional work must be completed.

The flexible, bivariate multistage clinical trial design searches over a multinomial distribution in a similar manner to the Conaway and Petroni (1995) methodology. Although the algorithm functions on a computer cluster, it will not execute in a reasonable time on a personal PC. It seems reasonable that the search algorithm can be improved through intelligent selection of the starting values. R is currently used to perform the search algorithm, but a precompiled language, such as C+ or Fortran, might be a better choice. If this problem is solved, it will also make the design more accessible.

Inference after the conclusion of single-arm, mutlistage, bivariate clinical trials also requires additional attention. Unbiased point estimation is an improvement over the maximum likely estimates, but additional research is required. Principal investigators and biostatisticians require confidence intervals, as well as p-values. The confidence intervals involve an evaluation of the coverage probabilities. The creation of the p-values requires an ordering of the sample space, which includes consideration of the toxicity monitoring.

Finally, the bivariate phase IIb or III clinical trial is only the first component of the research in this area. The initial design assumes that both endpoints are binomial, but the phase III clinical trial is usually designed to evaluate various survival rates. The current design, including the conditional power approach to interim analysis, can be applied in the situation that considers a survival probability and the binomial for toxicity. The design can also accommodate the same flexibility available in the phase II setting. The research should modify the multiple-arm, multiple-endpoint design so it can monitor toxicity on a different schedule than response. Finally, an unbiased estimator should be available in the multiple-stage, multiple-endpoint comparative clinical trial.

The proposed research represents future work that would improve the safety of the subjects while also ensuring the overall type I and II error rates are not sacrificed. The designs could also include adaptive features that, for example, estimate the correlation between response and toxicity based on available data; then the procedure modifies the sample size accordingly. The future research must also reflect the methods used in practice. The work will be directly influenced by current applications of statistical methods and may include considerations for patient heterogeneity based on emerging genetic methods.

REFERENCES

ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH Harmonized Tripartite Guideline. *Statistics in Medicine*, 18(15):1903–1942, 1999.

Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics. *Biotechnology Law Report*, 26(4):375–386, 2007.

National Cancer Institute: Common terminology criteria for adverse events v4.0, 2009. URL `http://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_8.5x11.pdf`.

A. C. Aitken and H. T. Gonin. On fourfold sampling with and without replacement. *Proceedings of the Royal Society of Edinburgh*, 55:114–125, 1935.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244, 1969.

L. A. Aroian. Sequential analysis, direct method. *Technometrics*, 10(1):125–132, 1968.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

J. Bryant and R. Day. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*, 51(4):1372–83, 1995.

M. Chang. Estimation of multiple response rates in phase II clinical trials with missing observations. *Journal of Biopharmaceutical Statistics*, 19(5):791–802, 2009.

S. C. Chow and J. P. Liu. *Design and Analysis of Clinical Trials: Concepts and Methdologies.* John Wiley & Sons, Inc., Hoboken, 2nd edition, 2004.

S. C. Chow, J. Shai, and W. Hansheng. *Sample Size Calculations In Clinical Research.* Chapman & Hall/CRC., Boca Rotan, 2nd edition, 2008.

C. Chuang-Stein, P. Stryszak, A. Dmitrienko, and W. Offen. Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine*, 26(6):1181–92, 2007.

T. Colton and K. McPherson. Two-stage plans compared with fixed-sample-size and wald sprt plans. *Journal of the American Statistical Association*, 71(353):80–86, 1976.

M. R. Conaway and G. R. Petroni. Bivariate sequential designs for phase II trials. *Biometrics*, 51 (2):656–664, 1995.

R. J. Cook and V. T. Farewell. Guidelines for monitoring efficacy and toxicity responses in clinical trials. *Biometrics*, 50(4):1146–52, 1994.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

J. Crowley and D. P. Ankerst. *Handbook of statistics in clinical oncology.* Chapman & Hall/CRC, Boca Raton, 2nd edition, 2006.

D. L. Demets, C. D. Furberg, and L. M. Friedman. *Data Monitoring in Clinical Trials: A Case Studies Approach.* Springer, 2005.

S. S. Emerson and T. R. Fleming. Symmetric group sequential test designs. *Biometrics*, 45(3): 905–23, 1989.

T. R. Fleming. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, 38 (1):143–151, 1982.

L. M. Friedman, C. D. Furberg, and D. L. DeMets. *Fundamentals of Clinical Trials.* Springer, New York, 3rd edition, 1998.

E. A. Gehan. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*, 13:346–53, 1961.

A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992.

A. Genz. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14:151–160, 2004.

M. Halperin, K. K. Lan, J. H. Ware, N. J. Johnson, and D. L. DeMets. An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials*, 3(4):311–23, 1982.

H. Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3): 360–378, 1931.

A. Ivanova, B. F. Qaqish, and M. J. Schell. Continuous toxicity monitoring in phase II trials in oncology. *Biometrics*, 61(2):540–5, 2005.

C. Jennison and B. W. Turnbull. Exact calculations for sequential t, chi-squared, and F tests. *Biometrika*, 78(1):133–141, 1991.

C. Jennison and B. W. Turnbull. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics*, 49(3):741–52, 1993.

C. Jennison and B. W. Turnbull. *Group Sequential Methods: Applications to Clinical Trials (Chapman & Hall/Crc Interdisciplinary Statistics Series)*. Chapman & Hall/CRC, 1999.

H. Jin. Alternative designs of phase II trials considering response and toxicity. *Contemporary Clinical Trials*, 28(4):525–31, 2007.

S. H. Jung and K. M. Kim. On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine*, 23(6):881–96, 2004.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

G. Kordzakhia, O. Siddiqui, and M. F. Huque. Method of balanced adjustment in testing co-primary endpoints. *Statistics in Medicine*, 29(19):2055–66, 2010.

J. M. Lachin. A review of methods for futility stopping based on conditional power. *Statistics in Medicine*, 24(18):2747–64, 2005.

K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

P. C. O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4): 1079–1087, 1984.

P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35 (3):549–56, 1979.

W. Offen, C. Chuang-Stein, A. Dmitrienko, G. Littman, J. Maca, L. Meyerson, R. Muirhead, P. Stryszak, A. Boddy, K. Chen, K. Copley-Merriman, W. Dere, S. Givens, D. Hall, D. Henry, J.D. Jackson, A. Krishen, T. Liu, S. Ryder, A.J. Sankoh, J. Wang, and C.H. Yeh. Multiple co-primary endpoints: medical and statistical solutions. a report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of america. *Drug Information Journal*, 41:31–46, 2007.

S. Pampallona and A. A. Tsiatis. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42(1-2):19–35, 1994.

S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

S. J. Pocock, N. L. Geller, and A. A. Tsiatis. The analysis of multiple endpoints in clinical trials. *Biometrics*, 43(3):487–498, 1987.

H. E. Ray and S. N. Rai. Operating characteristics of a Simon 2-Stage phase II clinical trial design incorporating continuous toxicity monitoring. *Under Revision, 2nd Round*, 2011a.

H. E. Ray and S. N. Rai. Flexible bivariate phase II clinical trial design incorporating toxicity and response on different schedules. *Under Review*, 2011b.

D. M. Reboussin, D. L. DeMets, KyungMann Kim, and K. K. Gordon Lan. ld98.exe, 1998.

J. R. Schultz, F. R. Nichol, G. L. Elfring, and S. D. Weed. Multiple-stage procedures for drug screening. *Biometrics*, 29(2):293–300, 1973.

R. Simon. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1): 1–10, 1989.

E. Slud and L. J. Wei. Two-sample repeated significance tests based on the modified wilcoxon statistic. *Journal of the American Statistical Association*, 77(380):862–868, 1982.

N. Stallard, J. Whitehead, S. Todd, and A. Whitehead. Stopping rules for phase II studies. *British Journal of Clinical Pharmacology*, 51(6):523–9, 2001. .

D. I. Tang, C. Gnecco, and N. L. Geller. Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association*, 84(407):776–779, 1989.

P. F. Thall, R. M. Simon, and E. H. Estey. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *Journal of Clinical Oncology*, 14(1):296–303, 1996.

C. Tournoux, Y. De Rycke, J. Medioni, and B. Asselain. Methods of joint evaluation of efficacy and toxicity in phase II clinical trials. *Contemporary Clinical Trials*, 28(4):514–24, 2007.

L. J. Wei, John Q. Su, and John M. Lachin. Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika*, 77(2):359–364, 1990.

P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing.* John Wiley & Sons, Inc., New York, 1993.

C. Wu and A. Liu. An adaptive approach for bivariate phase II clinical trial designs. *Contemporary Clinical Trials*, 28(4):482–6, 2007.

Z. Ying and W. R. Clarke. A flexible futility monitoring method with time-varying conditional power boundary. *Clinical Trials Journal*, 7(3):209–18, 2010.

L. Zhang and W. F. Rosenberger. Sequential monitoring of randomization tests. In Ravindra Khattree and Dayanand N. Naik, editors, *Computational methods in biomedical research*, Chapman & Hall/CRC biostatistics series. Chapman & Hall/CRC, Boca Raton, FL, 2008.

Y. Zhao, P. M. Grambsch, and J. D. Neaton. Comparison of numerical algorithms for bivariate sequential tests based on marginal criteria. *Computational Statistics & Data Analysis*, 49(3): 631–641, 2005.

CURRICULUM VITAE

NAME: Herman E Ray

ADDRESS: Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY 40292

EDUCATION: B.S. Mathematics
Middle Tennessee State University
2000

M.S. Mathematics
Middle Tennessee State University
2004

PROFESSIONAL POSITIONS: Research Scientist
Thomson Reuters, Chicago, IL
September, 2010 to Present

Senior Statistician
Thomson Reuters, Ann Arbor, MI
July, 2009 to September, 2010

Research Assistant
JG Brown Cancer Center
Biostatistics Shared Facility, Louisville, KY
August, 2009 to Present

Statistician
Thomson Reuters (Medstat), Nashville, TN
January, 2005 to July 2009

Freelance Consultant
Inclinix, Annapolis, MD
March, 2006 to March, 2008

Senior Statistician
CIGNA Medicare, Nashville, TN
May, 2001 to January, 2005

TEACHING:

Teaching Assistant
University of Louisville
Louisville, KY
Advanced Clinical Trials, Spring 2011

Adjunct Instructor
Nashville State Community College
Nashville TN
College Algebra, Fall 2001

HONORS AND AWARDS:

Golden Key Honor Society, 2010
Phi Kappa Phi Honor Society, 2004
Accomplishment Award, CIGNA Healthcare, 2003
Thomas Forrest Abstract Algebra Award, 2001

PROFESSIONAL AFFILIATIONS:

American Statistical Association

PUBLICATIONS

REVIEWED:

H.E. Ray and S.N. Rai. An Evaluation of a Simon 2-Stage Phase II Clinical Trial Design Incorporating Toxicity Monitoring. *Contemporary Clinical Trials*, May 2011, 32, 428 - 436.

IN PREPARATION:

H.E. Ray and S.N. Rai. Operating Characteristics of Simon 2-Stage Phase II Clinical Trial Design Incorporating Continuous Toxicity Monitoring. *Under Revision: 2nd Round*

H.E. Ray and S.N. Rai. Flexible Bivariate Phase II Clinical Trial Design incorporating Toxicity and Response on Different Schedules. *Under Review*

WHITE PAPERS:

L. MacCracken, G. Ray, and G. Popa. Optimizing Unique Market Characteristics, A Refined Process to Plan, Market, and Grow Your Outpatient Business. Thomson Reuters (2011).

G. Pickens, G. Popa, G. Ray. Thomson Reuters Healthcare Indexes: Consumer Confidence. Thomson Reuters (2009)

C. Kassed, D. Lewandowski, L. MacCracken, G. Pickens, G. Ray, L. Ray. Delayed Arrival: The Domestic Healthcare Traveler. Thomson Reuters (2008)

PRESENTATIONS:

L. MacCracken, G. Ray, G. Popa, and A. Skarulis (2011). Optimizing Unique Market Characteristics, A Refined Process to Plan, Market, and Grow Your Outpatient Business. Customer Web-based Seminar given for Thomson Reuters; Marketing and Planning Users Group; Ann Arbor, MI.

H.E. Ray, (2011). Inference in a Multistage Single-arm Trial Incorporating Multiple Endpoints. Seminar given at the School of Public Health, Department of Biostatistics and Bioinformatics, University of Louisville; Louisville, KY

H.E. Ray, (2010). Operating Characteristics of a Simon 2-Stage Phase II Clinical trial Design Incorporating Continuous Toxicity Monitoring. Seminar given for Thomson Reuters, Applied Analytics; Ann Arbor, MI.

H.E. Ray, (2010). Introduction to R. Seminar given for Thomson Reuters, Applied Analytics; Ann Arbor, MI.

H.E. Ray, (2010). Operating Characteristics of a Simon 2-Stage Phase II Clinical trial Design Incorporating Continuous Toxicity Monitoring. Seminar given at the Joint Statistical Conference; Vancouver, British Columbia.

H.E. Ray and B. Hochrein (2010) Point Estimation and Standard Deviation, Statistical Education Series. Thomson Reuters, Analytic Services; Chicago, IL.

H.E. Ray, (2009). Introduction to SAS Graph. Seminar given for Thomson Reuters, Applied Analytics; Chicago, IL.

H.E. Ray, (2007). SAS Connect. Seminar given for Thomson Reuters, SAS Users Group; Chicago, IL.

H.E. Ray, (2004). Data Driven Approach to Prioritizing Work. Seminar given at the 7th Annual Conference on Statistics in the Medicare Program; La Guardia, NY.